UNSUPERVISED AND SEMI-SUPERVISED TRAINING METHODS FOR EUKARYOTIC GENE PREDICTION

A Dissertation
Presented to
The Academic Faculty

By

Vardges Ter-Hovhannisyan

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics

Georgia Institute of Technology

December, 2008

UNSUPERVISED AND SEMI-SUPERVISED TRAINING METHODS FOR EUKARYOTIC GENE PREDICTION

Approved by:

Dr. Mark Borodovsky, Advisor
Department of Biomedical Engineering and Computational Science and Engineering Division
Georgia Institute of Technology

Dr. Leonid Bunimovich
School of Mathematics
Georgia Institute of Technology

Dr. Jung H. Choi
Department of Biology
Georgia Institute of Technology

Dr. Yury Chernoff
Department of Biology
Georgia Institute of Technology

Dr. King Jordan
Department of Biology
Georgia Institute of Technology

Date Approved November 6, 2008

To my parents, Rima and Edmund,
my brother, Aleksandr,
my wife, Maga, and my son, Edward

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Mark Borodovsky, for the opportunity and the guidance. Special thanks to Aleksandre Lomsadze for motivation, support, and friendship, and my wife, Maga Khachatryan, for her patience and her help with my first steps in statistics and computer programming.

I thank Natalya Shmeleva, Yuan Tian, and Andrey Kislyuk for their work in developing test sets for algorithm performance evaluation. I extend my thanks to John Besemer for useful discussions and comments.

I would also like to express my gratitude to Dr. Yury Chernoff for elucidation of biological significance of newly identified genes and my committee members Dr. Leonid Bunimovich, Dr. Jung Choi, and Dr. King Jordan for their time and comments.

I am very grateful to Aram Partizian and Maga for assistance in proof-reading the manuscript.

This work was supported by grant HG00783 from the US National Institutes of Health (NIH) and by the School of Biology at Georgia Institute of Technology.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

3' UTR – 3' untranslated region

5' UTR – 5' untranslated region

aa – Amino acid

A,T,G,C – The nucleotides Adenosine, Guanosine, Tymidine, Cytidine

blast – Basic local alignment search tool

BP – Branch point

CDD – Conserved Domain Database

cDNA – Complementary DNA

COG – Clusters of Orthologous Groups of proteins (COGs)

CPU – Central Processing Unit

DOE – Department of Energy

EM – Expectation Maximization

EST – Expressed sequence tag

G+C% – Percent of C and C nucleotides

GFF – General Feature Format

GHz – Gigahertz

HMM – Hidden Markov Model

JGI – Joint Genome Institute

Kb – Kilobases

*KL d*istance – Kullback-Liebler distance

Mb – Megabases

MHz – Megahertz

MIPS – Munich Information Center for Protein Sequences

NCBI – National Center for Biotechnology Information

*NR dat*abase – Non redundant database

nt – Nucleotide

*OR*F – Open Reading Frame

P – Probability

Sn – Sensitivity

Sp – Specificity

*SM*P – Symmetric multiprocessing

*TBF*5 – Transcription factor complex subunit Tfb5

SUMMARY

This thesis describes new gene finding methods for eukaryotic gene prediction. The current methods for deriving model parameters for gene prediction algorithms are based on curated or experimentally validated set of genes or gene elements. These training sets often require time and additional expert efforts especially for the species that are in the initial stages of genome sequencing. Unsupervised training allows determination of model parameters from anonymous genomic sequence with. The importance and the practical applicability of the unsupervised training is critical for ever growing rate of eukaryotic genome sequencing.

Three distinct training procedures are developed for diverse group of eukaryotic species. GeneMark-ES is developed for species with strong donor and acceptor site signals such as *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Drosophila melanogaster*. The second version of the algorithm, GeneMark-ES-2, introduces enhanced intron model to better describe the gene structure of fungal species which posses with relatively weak donor and acceptor splice sites and well conserved branch point signal. GeneMark-LE, semi-supervised training approach is designed for eukaryotic species with small number of introns.

The results indicate that the unsupervised training methods perform well as compared to other training methods and as estimated from the set of genes supported by EST-to-genome alignments. The analysis of novel genomes led to interesting biological findings and showed that several of fungal species are either over-annotated or under-annotated.

CHAPTER 1
INTRODUCTION

1.1 Purpose and Scope of Research

At the dawn of the twenty first century genome sequencing of several model organisms has created a stage for new research paradigms, transforming biology into a data-driven science. Recent technological and scientific achievements, including DNA sequencing technology (Ghadessy, Ong et al. 2001, Dressman, Yan et al. 2003, Margulies, Egholm et al. 2005), have provided a unique opportunity to decrease the cost, time and complexity involved in deciphering large DNA molecules. Interdisciplinary approaches have become very successful in the post-genomics era; computer algorithms and mathematics now play significant role in biological sciences.

The DNA sequence data is most valuable to the biological community when it is supplied along with quality annotation. Fundamental steps after genome sequencing include transformation of the cryptic language of life into meaningful knowledge about the genes, protein structure and function that these genes encode, identification of regulatory elements, and illumination of the evolutionary history. Although experimentally found genes are the most reliable annotations, experiments are time-consuming and costly. Currently manual annotation - the 'gold standard'- is the outcome of diligent analysis and manual verification that rely on multiple sources of evidence.

At the time that this thesis was initiated, sequencing of only six eukaryotic and sixty prokaryotic genomes had been completed. At the beginning of 2008 these numbers had increased more than five-fold, and the number of eukaryotic genome projects in

progress exceeded thousand (http://www.genomesonline.org/ and Liolios, Tavernarakis et al. 2006). While the rate at which sequences submitted to GenBank continues to grow exponentially, the number of researchers that can provide valuable expertise for proper annotation is insufficient. Computer programs generally assist experts in the fulfillment of their goals since the annotation process from scratch by human without such computational power in principle would be impossible.

Although many gene prediction algorithms have been developed over the years the annotation is hardly keeping pace with the supply of raw sequence data. Particularly, in the case of eukaryotic genome annotation, the process is slowed down by the underlying structure of the most gene finders which require either a set of validated genes to derive model parameters for *ab initio* gene finder or representative databases of ESTs and/or cDNAs or genomic sequences of closely related species for extrinsic approaches.

For prokaryotic species this problem has been addressed by several unsupervised training methods (Audic and Claverie 1998, Frishman 1998, Hayes and Borodovsky 1998, Besemer, Lomsadze et al. 2001, Larsen and Krogh 2003). Self-training algorithms for eukaryotic species was assumed to be unfeasible task. The research efforts of this work are focused on model parameter estimation for gene finding in eukaryotic genomes from anonymous genomic DNA without a training set. This thesis consists of three main parts with a core that describes a novel unsupervised training method for eukaryotic gene prediction. The scope of this research is extremely important considering (i) the growing number of eukaryotic species that undergo large scale sequencing but lack the data necessary for application of traditional gene finding methods and (ii) the time and efforts invested into building reliable training sets.

Two versions of unsupervised iterative training procedures are described and applied to eukaryotic species that vary in their gene organization and genome size. The first, GeneMark.hmm ES 3.0 (E-eukarytotic, S- self-training, 3.0- GeneMark.hmm version) is used to derive model parameters for species possessing introns with strong signals around the splice sites (Lomsadze, Ter-Hovhannisyan et al. 2005) such as *Arabidopsis thaliana* (Initiative 2000), *Caenorhabditis elegans* (Consortium 1998) and *Drosophila melanogaster* (Adams Celniker et al. 2000). For convenience in what follows GeneMark.hmm ES 3.0 will be referred as GeneMark-ES.

The second, extended version of GeneMark-ES (GeneMark-ES-2 in what follows), employs a new enhanced intron submodel (Ter-Hovhannisyan, Lomsadze et al. 2008). It is better suitable for species which contain a significant part of the information for intron splicing in the branch point (BP) site (typical for most of the fungal species).

Spliced gene element parameterization via both unsupervised and supervised training methods is a challenging task in low eukaryotes. The difficulty is associated with the limited size of data reflecting structure of the spliced gene. Semi-supervised gene finding approach, GeneMark-LE, is developed for low eukaryotes which contain relatively small number of introns. The algorithm utilizes information reflecting multiple exon gene structure e.g. donor, acceptor and BP signals, available from closely related species in combination with unsupervised training to derive species-specific model parameters for gene identification.

The performance of the algorithms is tested on validated gene sets and compared to traditional (supervised) *ab initio* training approaches. The results show that GeneMark-ES performs as well as or better that other *ab initio* supervised gene finding

methods. Genes predicted on whole genome level are compared to the publicly available sets of annotated genes to identify novel genes.

Introduction of the new intron submodel in GeneMark-ES-2 leads to a significant increase in gene prediction accuracy for fungal genomes. Predictions of this approach are also compared to annotation. Interestingly, for most of the fungal species the GeneMark-ES-2 generates results similar to annotation which is based on several gene finding methods, often manually curated and produced in a course of several years. GeneMark-LE is shown to perform with high levels of gene prediction outperforming gene models that are derived from data extracted from closely related species.

GeneMark-ES and GeneMark-ES-2 are currently part of the annotation process used by several sequencing and annotation centers such as the Joint Genome Institute (JGI), the Broad Institute, the Munich Information Center for Protein Sequences (MIPS) and WormBase. In fact, a number of requests for GeneMark-ES (particularly from the institutions listed above) and encouraging user feedback led to its further development to GeneMark-ES-2.

1.2 General Overview of Species

The species under study are of significant scientific interest as they can be harmful and/or beneficial organisms for humans. For example, the sequence and annotation of *Anopheles gambiae* (Holt Subramanian et al. 2002), the principle vector of malaria, is believed to be a valuable resource in the prevention and treatment of malaria outbreaks (Holt Subramanian et al. 2002). *Toxoplasma gondii*, a protozoan parasite, causes infection in warm-blooded animals (Boothroyd and Grigg 2002). A wide variety

of fungi attack plants, causing devastating crop infections. *Fusarium* species are plant pathogens that cause various diseases to nearly all economically important plant species. More than four hundred plant species are hosts to the fungal plant pathogen *Sclerotinia sclerotiorum* (http://www.ncbi.nlm.nih.gov/).

Although the overwhelming number of fungal genomes are plant pathogens and are important to agriculture and industry, they also cause diseases in humans. The encapsulated fungal organism *Cryptococcus neoformans* (Loftus, Fung et al. 2005) affects individuals with a weak immune system and may lead to serious complications including fatal meningitis (Mitchell and Perfect 1995). Valley fever, a respiratory infection, is caused by *Coccidioides immitis* and is acquired by humans from inhalation of infected spores (Drutz and Catanzaro 1978). Complications from infection caused by *Histoplasma capsulatum* include inflammation of pericardium and fibrosis of blood vessels (Durkin, Kohler et al. 2001). Fungi are also successfully utilized in food and pharmaceutical industries. *Aspergillus niger* (Pel, de Winde et al. 2007), responsible for one of the most efficient bioprocesses, is used as an enzyme in the fermentation industry for production of citric acids and represents one of the most efficient bioprocesses.

1.3 Program Availability

Both programs *GeneMark-ES-3.0* and GeneMark-ES-2 are publicly available for download at http://topaz.gatech.edu/GeneMark/GeneMark-ES-2008. The models already derived by GeneMark-ES are available for use at http://topaz.gatech.edu/GeneMark/eukhmm.cgi. The GeneMark-LE package is currently under development. The program should be publicly available by the end of the 2008.

1.4 Additional Supplementary Materials

Test sets used for algorithm performance evaluation are available http://exon.gatech.edu/GeneMark/GeneMark-ES-2008/TEST_SETS.tar.gz. The predictions of GeneMark-ES for novel genes are available at http://nar.oxfordjournals.org/cgi/content/full/33/20/6494/DC1. Full list of novel genes predicted by GeneMark-ES-2 in fungal species are available at http://exon.gatech.edu/GeneMark/GeneMark-ES-2008/Novel_genes.

1.5 Thesis Outline

The rest of the thesis chapters are organized as follows. *Chapter 2* presents an overview of the relevant research in the area of gene finding. *Chapter 3* describes materials and general methods used to conduct this research. *Chapter 4* introduces and discusses the results of GeneMark-ES, a self-training algorithm for eukaryotic gene finding. *Chapter 5* describes the gene finding in fungal genomes via unsupervised training utilized by GeneMark-ES-2. GeneMark-LE, a gene finding algorithm for genomes with low numbers of introns, is described in *Chapter 6*. *Chapter 7* discusses the current challenges and the future directions for unsupervised training.

CHAPTER 2
LITERATURE REVIEW

Gene finders usually consist of the prediction algorithm and the models that describe genome-specific characteristics of a species in hand. The algorithm reflecting the mathematical structure of the predictor uses features provided by the models to find gene(s) in a given genomic sequence. The design of the algorithm usually requires *a priori* knowledge about the genome organization (e.g. prokaryotic, eukaryotic). The models depending on the algorithm framework, particularly for *ab initio* gene prediction, can vary significantly.

Studies on protein coding gene prediction began in the early 1980s (Shepherd 1981, Fickett 1982). Since then many gene finding algorithms have been developed. Originally, two distinct approaches for gene identification were developed: (i) intrinsic (Borodovsky and McIninch 1993, Kulp, Haussler et al. 1996, Burge and Karlin 1997, Krogh 1997, Lukashin and Borodovsky 1998, Delcher, Harmon et al. 1999, Salzberg, Pertea et al. 1999, Parra, Blanco et al. 2000, Reese, Kulp et al. 2000, Salamov and Solovyev 2000, Majoros, Pertea et al. 2003, Stanke and Waack 2003, Korf 2004, Lomsadze, Ter-Hovhannisyan et al. 2005) that uses statistical models to describe gene elements, and (ii) extrinsic (Gish and States 1993, Gelfand, Mironov et al. 1996, Rogozin, Milanesi et al. 1996, Kulp, Haussler et al. 1997, Florea, Hartzell et al. 1998, Laub and Smith 1998, Mironov, Roytberg et al. 1998, Badger and Olsen 1999, Delcher, Kasif et al. 1999, Pachter, Batzoglou et al. 1999, Bafna and Huson 2000, Batzoglou,

Pachter et al. 2000, Kent and Zahler 2000, Morgenstern 2000, Schwartz, Zhang et al. 2000, Kent 2002, Meyer and Durbin 2002, Morgenstern, Rinner et al. 2002, Pachter, Alexandersson et al. 2002, Schlueter, Dong et al. 2003, Schwartz, Elnitski et al. 2003, Birney, Clamp et al. 2004) that utilizes comparative methods based on the information obtained from a database (e.g. cDNA library, protein database). Each of these techniques has its strengths and weaknesses. While the intrinsic (*ab initio)* methods exhibit high sensitivity, the main criticism is their rather high rate of false positive predictions and their dependence on a reliable and representative training set. In contrast, extrinsic methods are highly specific but are limited by a number of factors including the database size and quality as well as the difficulty to obtain gene structure even in the presence of good similarity to the entry in the database. A comprehensive review of the advantages and weaknesses of these methods is done by Mathe et al. (Mathe, Sagot et al. 2002).

Algorithms that utilize Markov chain theory to describe protein-coding regions in DNA sequence became popular more than twenty years ago (Borodovsky 1986a, Borodovsky 1986b, Tavare and Song 1989). The first gene identification algorithm that used inhomogeneous and homogenous Markov models describing coding and non-coding regions respectively, GeneMark (Borodovsky and McIninch 1993), uses Bayesian analysis to estimate the *a posteriori* probability of a fragment in a given DNA sequence to be either coding or non-coding region.

Markov models were integrated into the framework of Hidden Markov Models (HMM), statistical or probabilistic models initially utilized in speech recognition (Rabiner 1989). The HMM was introduced to gene finding in the mid-1990s (Krogh, Brown et al. 1994). The major drawback of classical HMM is the geometrical state

duration, which is unsuitable for the biological systems. Modeling durations explicitly permit length distributions of a desirable shape (Rabiner 1989). Such a modification of HMM, called hidden semi-Markov model (HSMM) (also cited in literature as HMM with duration, or generalized HMM), is used in number of gene finders (Parra, Blanco et al. 2000, Reese, Kulp et al. 2000, Salamov and Solovyev 2000, Majoros, Pertea et al. 2003, Stanke and Waack 2003, Korf 2004, Lomsadze, Ter-Hovhannisyan et al. 2005). Reportedly, HMM-based gene finders demonstrated the best performance in the Genome Annotation Assessment Project (GASP) (Reese 2000).

GENSCAN (Burge and Karlin 1997), an *ab initio* eukaryotic gene finder that was used in annotation of the human genome, introduced HSMM (independently also described by Kulp *et al.* (Kulp, Haussler et al. 1996)). A statistical decomposition model was used to determine the most informative splice sites. A windowed weight array matrix (WWAM) of second order is used to describe the BP site. GENSCAN provides gene prediction models for the human genome (and often for other vertebrates as well), *A. thaliana* and maize. The underlying HMM architecture used by GENSCAN describes nearly all known gene elements and contains states for the multiple and the single exon genes, the non-coding region, splice sites, the BP site, the translation initiation and termination sites, and the promoter and polyadenylation signals. The wealth of literature on other gene features such as splicing regulatory elements is increasing (Fairbrother, Yeh et al. 2002, Wang and Burge 2008). GENSCAN also includes the *a posteriori* probability of a predicted coding region by the Viterbi decoding in the program output.

HMM-based *ab initio* gene finders such as FGENESH (Salamov and Solovyev 2000), SNAP (Korf 2004), GeneMark.hmm (Lomsadze, Ter-Hovhannisyan et al. 2005)

and AUGUSTUS (Stanke and Waack 2003) were developed. Splice site detection in short introns was addressed in INTRONSCAN (Lim and Burge 2001). Log-odds scores that are assigned to splice signals, a BP site and intron length combined with composition scores contribute to the total score of the potential intron.

AUGUSTUS uses an intron submodel to describe short and long introns as mixture distributions based on the fact that the splicing mechanism undergoes either by the process of intron definition which states that the splicing machinery recognizes and pairs the splice sites across the intron (short introns and long exons) or by an opposite process of exon definition model according to which splice sites are identified and paired across the exon (Berget 1995). The authors used a training set to model the short intron length distribution and explicitly defined an exponential distribution to describe the long introns. A practical but time-consuming feature integrated within the AUGUSTUS package is that it allows the user to optimize model parameters given the training set. This option also provides an opportunity to update model parameters when better and/or additional data becomes available. The BP site is modeled similarly to GENSCAN except that a third order WWAM emitting a 32nt long sequence is used instead (Stanke and Waack 2003). Models describing the BP are utilized in other supervised gene finding algorithms such as NetGene2 (Hebsgaard, Korning et al. 1996) and GipsyGene (Neverov 2003).

The parameters of the BP site were derived from alignment of putative BP sites by simulated annealing (Lukashin, Engelbrecht et al. 1992), the Expectation Maximization (EM) algorithm (Hebsgaard, Korning et al. 1996) and the Markov Chain Monte Carlo (MCMC) method (Lim and Burge 2001). Hebsgaard *et al.* modified the

sampling procedure by locking on a predefined BP, the nucleotide "A" (Hebsgaard, Korning et al. 1996). Neverov *et al.* created initial branch point site profiles based on information derived from related species (Neverov 2003). Lim *et al.* decreased the search space by removing motifs which did not contain consensus the BP site (Lim and Burge 2001).

SNAP (Korf 2004) has an HMM architecture similar to GENSCAN's but with reduced HMM state space. Several states including hidden states for promoters, polyadenylation signals and UTRs are not part of SNAP's HMM architecture. In addition SNAP's HMM diagram allows the user to vary the length of the weight matrix and the order of the Markov model to better fit a particular genome. For novel genomes it uses a "bootstrap" algorithm to derive model parameters for a given species. Although the algorithm performance with models derived by the bootstrap approach was reported to be lower than that of models derived by supervised training, the accuracy of gene prediction was within a reasonable range of values.

Multi-step gene prediction approaches such as GeneID (Guigo, Knudsen et al. 1992, Parra, Blanco et al. 2000) and mGene (previously named G3A http://www.fml.tuebingen.mpg.de/ raetsch/projects/mgene2) were also developed. GeneID hierarchically progresses from sites to exons to genes. First it predicts potential sites such as gene start, stop and splice sites using Positional Weight Matrices (PWM). Then it assigns the log-odds scores to exons as the sum of the PWM scores and the log-odds ratios of a Markov Model for coding DNA. At the final step the dynamic programming algorithm (Guigo 1998) is applied to determine the best gene structure. mGene is a two step approach which consists of "layers". In the first layer Support

Vector Machines (SVM) (Sonnenburg 2005) create the initial seeds of signal sites and other genomic features (e.g. intergenic region, coding, intron). To determine the best gene structures these seeds are analyzed by the second layer which employs Semi Hidden Markov SVMs (Raetsch 2005).

Conditional random fields (CRFs), a framework for probabilistic modes, has became popular in recent years (Lafferty 2001). An attractive feature of this technique is the scoring function which takes into account a segment of the past k occurrences consequently relaxing the assumption of independence in the theory of HMM; this additional option, however, significantly increases the complexity of calculations. Bernal et al. used CRFs in the gene finding tool CRAIG (Bernal, Crammer et al. 2007). This discriminative training method estimates all model parameters simultaneously (assuming intra-state dependency) to maximize annotation accuracy. While the empirical threshold was used to classify introns into short or long categories weights were applied to "globally balance" model components. Analogous to how the HMM adapted to HSMM to better reflect the state durations, CRFs were modified to better model the segment properties (e.g. segment length, transition within segment) in semi-CRFs (Sarawagi 2004).

Extrinsic methods generally acquire additional information supplied from an external source such as cDNAs, EST or protein databases (Gish and States 1993, Rogozin, Milanesi et al. 1996, Mironov, Roytberg et al. 1998). Gene prediction accuracy can be significantly improved given a reliable cDNA expressed by a particular DNA sequence. This, however, is an ideal scenario and the reality is more complicated. One of the disadvantages of the extrinsic approaches is that they are essentially database-dependent and may fall short of providing sufficient support for gene annotation in novel

genomes. In addition the sets of available ESTs and cDNAs are frequently related to highly expressed genes which could lead to miss-annotation of low expressed genes (22). EST contamination, on the other hand is a problem caused by the type of procedure that is used to construct a cDNA library (Bonaldo, Lennon et al. 1996, Krizman, Chuaqui et al. 1996, Peterson, Brown et al. 1998, Camargo, Samaia et al. 2001) and typically involves sequences from parts of the sequencing linker or vector flanked at the EST's ends or vector contamination from DNA rearrangement inside the bacteria leading to the inclusion of foreign sequences into the EST. Reconstruction of a complete gene structure by alignment methods using ESTs that partially cover mRNAs is not a challenging task. As the wealth of information grows these type of extrinsic methods become more popular (Birney, Clamp et al. 2004).

Several *ab initio* gene finders were extended to use extrinsic information. HMMGene (Krogh 2000) developed for the human and the worm "locks" the annotation supplied by the user and predicts the gene structure given the constrains. GenomeScan (Yeh, Lim et al. 2001) employs user-defined protein homology information to enhance GENSCAN's gene predictions by assigning bonus values to regions consistent with extrinsic information, which is derived from protein and or EST/cDNA. The authors introduce heuristic setting *root-r* to account for false positive hits produced by similarity search, *blastx* (Gish and States 1993). GeneID (Parra, Blanco et al. 2000) ExonHunter (Brejova, Brown et al. 2005) and AUGUSTUS+ (Stanke, Schoffmann et al. 2006) combine extrinsic evidence, defined in terms of gene elements, with intrinsic approaches to recover the complete gene structure. To address the problem of the reliability of extrinsic evidence, GeneID provides an option of rating the extrinsic information with a

cumulative scoring function to reflect its reliability (http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml#cum_score_array). ExonHunter uses a more complicated scenario in which it combines different sources of extrinsic information (advisors) into a superadvisor and then determines such a gene structure by maximizing *P(gene|sequence, evidence)*. AUGUSTUS+ requires that the supplied "hints" are biologically relevant, e.g. there is no in-frame stop codon within a coding sequence or all the splice sites are canonical.

Dictionary based approaches were developed to obtain segments from a database to find matching fragments in the query sequence (Laub and Smith 1998, Pachter, Batzoglou et al. 1999). INFO (Laub and Smith 1998) employs a protein database search in six translation frames of query (DNA) sequence and then identifies splice sites around the conserved regions and later Patcher et al. utilized both EST and protein database in the search for the best gene structures (Pachter, Batzoglou et al. 1999).

EuGene (Schiex 2001) and GAZE (Howe, Chothia et al. 2002) use a dynamic programming algorithm to integrate intrinsic and extrinsic information into the eukaryotic gene prediction. GAZE, specifically designed for *C. elegans*, works indirectly with DNA sequences through an input file that contains the arbitrary signal and content sensors which is assembled into a complete gene structure using a configuration file. In addition, GAZE allows incorporation of trans-splicing (Conrad, Lea et al. 1995) into the external information. EuGeneHom (Foissac 2003), a gene finding approach tuned for plants, uses EuGene as a gene predictor and tblastx as a tool for conducting a homology search to predict genes in eukaryotic genomes. The program is available for on-line use with a 400

Kb limit for an input sequence (http://bioinfo.genotoul.fr/apps/eugene/ EuGeneHom/cgi-bin/EuGeneHom.pl).

GeneWise (Birney, Clamp et al. 2004), an evidence-based gene finding method for the human and the worm intensively used as part of Ensembl (Curwen, Eyras et al. 2004), uses a merged HMM model that reflects the alignment status between the protein database and the genomic DNA. The HMM, however, is rather simplistic: it accounts for direct strand only and does not consider intron phases other than zero.

ROSETTA (Batzoglou, Pachter et al. 2000) was developed for human and mouse genomes. It identifies synthetic blocks between pairs of homologous sequences by (i) employing a sequence alignment tool GLASS (Global Alignment SyStem) specifically designed to handle long (few hundred Kb) sequences and (ii) applying gene models similar to one featured in GENESCAN to find genes in the conserved regions (Burge and Karlin 1997). Currently direct and reverse strands are treated (by ROZETTA) separately. Two alignment techniques are integrated into Alignment-based Gene-Detection Algorithm, or AGenDA, (Rinner and Morgenstern 2002), a gene finding algorithm for primates and rodents. First AGenDA runs an alignment program, CHAOS (Brudno, Chapman et al. 2003), which creates the initial alignment map to reduce the search space, and then it runs the second aligner DIALIGN (Morgenstern, Dress et al. 1996) which implements local alignment of high homology regions identified by CHAOS. Finally, the complete gene structure is constructed frorm the potential splice sites.

Gene predictions in genome pairs were implemented in a number of programs. Pair HMM (Durbin 1998, Kent and Zahler 2000) and series of local alignments are used in Doublescan (Meyer and Durbin 2002) to reconstruct the gene structures. The HMM

architecture does not include length distributions for coding and non-coding states as well as promoter signals. This reduced HMM structure is justified by the sequence conservation of protein coding sequences within two species. The authors report better specificity results by applying a post-processing step which removes predictions that exhibit intron lengths less that 50 nt and/or a coding sequence (CDS) with length less than 120 nt.

SLAM (Alexandersson, Cawley et al. 2003) utilizes generalized pair HMM (GPHMM) which is based on combination of pair HMM and generalized HMM. To reduce the search space for GPHMM the algorithm concurrently builds an alignment of homologous regions in two genomes by using an external alignment tool, AVID (Bray, Dubchak et al. 2003). Then GPHMM is applied to find genes in the aligned segments of both genomes. The important question for the aligning step is: "What is the optimal divergence between two species?" In cases when two genomes are highly related, the alignment step might be somewhat redundant given the high conservation of non-coding regions. On the other hand conservation in protein coding regions becomes a problem in diverged species.

An informant genome approach has been developed for SGP2 (Parra, Agarwal et al. 2003) and TWINSCAN (Korf, Flicek et al. 2001). While both of these programs use an *ab initio* gene finder, GeneID (SGP2) and GeneScan (TWINSCAN), they employ extrinsic methods TBlastx (SGP2) and RepeatMasker (TWINSCAN) from a well-studied genome. TWINSCAN derives model parameters for the conserved regions from the training set. As a result the algorithm handles the conservation in protein coding and non-coding regions by different probabilistic models.

Comparative gene finding was extended to multi-genome gene finders which handle multiple sequence alignments; phylogenetic HMMs were introduced (Pedersen and Hein 2003, Siepel and Haussler 2004). These algorithms use HMM techniques often used in molecular evolution. Gene finder phylo-HMM (Siepel and Haussler 2004) can be adjusted for a single sequence gene finder when only one sequence is available, or it can be used as a gene finder in pair genome utilizing GPHMM when pairwise alignment is provided, or it can be run as phylo-HMM when multiple sequence alignment and subsequently the evolutionary tree is available. ExoniPhy (Siepel and Haussler 2004) identifies exons conserved in all domains of life by using (i) context-dependent phylogenetic models, (ii) explicit models for of conserved non-coding DNA, and (iii) models of insertions and deletions (indels). It finds the exons, not the genes, because the individual exons are more likely to be preserved throughout the evolution than are complete genes. The predicted exons can later be assembled into complete genes by a dynamic programming algorithm.

Whole genomic alignments of closely related species were used in several algorithms to find functionally important regions (Delcher, Kasif et al. 1999, Bafna and Huson 2000, Batzoglou, Pachter et al. 2000, Kent and Zahler 2000, Morgenstern 2000, Schwartz, Zhang et al. 2000, Schwartz, Elnitski et al. 2003). MUMmer (Delcher, Kasif et al. 1999) was initially designed for alignment of closely related prokaryotic genomes and later extended to eukaryotic species (Delcher, Phillippy et al. 2002, Kurtz, Phillippy et al. 2004). WABA (Kent and Zahler 2000) was used to align closely related nematode genomes of *Caenorhabditis briggsae* and well-studied *C. elegans* to identify conserved

regulatory elements. WABA proceeds with tree-step procedure employing seven-state pair-HMM for detailed alignment phase.

The Pairagon+N-Scan_EST (Arumugam, Wei et al. 2006) package includes two distinct algorithms for cDNA-to-genome alignment (Pairagon), and *de novo* predictor (N-Scan_EST; essentially modified version of TWINSCAN) which can take hints from ESTs. The algorithm uses native ESTs (those obtained from ESTs and/or cDNAs that are produced by the species under study) as an input for N-Scan_EST and exploits the evolutionary conservation between genome in hand and other mammalian organisms to analyze the alignments. The advantage of using native alignments is that there is no need to account for evolutionary divergence and the sequence similarity is nearly 100% between the DNA and its ESTs; the possible differences are caused by phenotypic/allelic variations or sequencing problems. It constructs states of so-called ESTseq from native EST alignments and labels it as (i) **I**, if it falls in the intron region of all overlapping EST alignments, (ii) **E**, if it falls in the exon region of all overlapping EST alignments, and (iii) **N**, if there is a disagreement in overlapping EST alignments and/or there are no overlapping EST alignments. HMM predictor (N-SCAN) emits ESTseq symbols together with target genome bases and conservation sequence symbols with parameters estimated from known gene structures and their ESTseqs. Homogeneous Markov chains for UTR, intron and coding states and a weight array model for splice sites are used to model ESTseq. While the authors have shown superior results for the human and the worm they reported difficulties in genes containing introns in UTR region.

Parra et al. combined several gene finding methods (GeneWise (Birney, Clamp et al. 2004), GeneID (Parra, Blanco et al. 2000)) and similarity searches, e.g. t-coffee

(Notredame and Suhre 2004), tblastn (Gertz, Yu et al. 2006), to produce a set of core genes in eukaryotic genomes (Parra, Bradnam et al. 2007).

Methods that combine the output of several predictors were developed recently. Posterior probability scores were used to combine outputs of two eukaryotic gene finding algorithms, GENSCAN and HMMgene (Rogic, Ouellette et al. 2002). Combiner (Allen, Pertea et al. 2004), a eukaryotic gene finder, utilizes several eukaryotic gene prediction programs (GENSCAN (Burge and Karlin 1997), GlimmerM (Salzberg, Pertea et al. 1999) GeneMark.hmm (Lomsadze, Ter-Hovhannisyan et al. 2005), GeneSplicer (Pertea, Lin et al. 2001), and TWINSCAN) in combination with sequence (EST, cDNA to genome) alignments. The most recent eukaryotic gene predictor, EVIGAN (Liu, Mackey et al. 2008), produces consensus gene models by combining various sources of evidence such as predictions of other gene finders, EST matches and protein-to-genome alignments. It uses EM algorithm to derive model parameters. Dynamic Bayesian Network is applied to predict genes by inferring the most likely consensus gene models given the evidence.

A gene annotation process carried by large sequencing centers such as Sanger (Ashurst, Chen et al. 2005), the J. Craig Venter Institute (http://www.jcvi.org/) and the Broad Institute (www.broad.mit.edu) takes into account all sources of information and employs a combination of intrinsic and extrinsic gene finding methods. These integrated approaches utilize similarity searches against protein and/or EST/cDNA databases. Ensembl's automated system annotates genes based on evidence from known protein, cDNA and EST data which is integrated into a MySQL database and the Perl Application Programming Interface (API). Sanger's The Vertebrate Genome Annotation (Vega)

provides resources for browsing manual annotation of finished sequences of vertebrate genomes. It too uses an Ensembl MySQL database to store the data and Ensembl pipeline to analyze it. The difference between Vega and Ensembl is that the former presents annotations coming from time-consuming process of manual curation, whereas the latter shows computationally derived gene predictions. Manual annotations for Vega pipeline are based on supporting transcriptional evidence. Set of rules is applied for identification of known, novel and putative genes as well as novel transcripts and pseudogenes (Ashurst, Chen et al. 2005).

Wealth of literature on eukaryotic gene prediction is overwhelming. Various gene finding methods are available for biological interpretation of genomic sequence. However, there is still a real need for accurate and fast gene prediction tools especially in initial stages of genome sequencing (Mathe, Sagot et al. 2002, Guigo, Flicek et al. 2006).

CHAPTER 3

MATERIALS AND GENERAL METHODS

3.1 Datasets for Unsupervised Training and Test Set Derivation

Genomes representing different taxonomic groups were used in this study. EST and genomic data used for unsupervised training and test set preparation was downloaded from GenBank (Benson, Karsch-Mizrachi et al. 2008), the Broad institute (www.broad.mit.edu), wormbase (www.worbase.org), flybase (www.flybase.org), the Joint Genome Institute (www.jgi.doe.gov), ToxoDB (toxodb.org) and Munich Information Center for Protein Sequences (MIPS). At the time this research was initiated in 2002, five species presented here- *A. thaliana* (Initiative 2000), *C. elegans* (Consortium 1998), *Drosophila melanogaster* (Adams Celniker et al. 2000), and *Schizosaccharomyces pombe* (Wood Gwilliam et al. 2002) - were well-studied model organisms with completed and published genome projects. Draft genomic assembly was publicly available for *Anopheles gambiae* (Holt Subramanian et al. 2002), *Aspergillus nidulans* (Galagan, Calvo et al. 2005), *Ciona intestinalis* (Dehal, Satou et al. 2002) and *Neurospora crassa* (Galagan, Calvo et al. 2003); other species such as *A. niger* (Pel, de Winde et al. 2007), *Aspergillus terreus*, *Botrytis cinerea*, *C. neoformans* (Loftus, Fung et al. 2005), *Chlamydomonas reindhardii*, (Merchant Prochnik et al. 2007) *C. immitis*, *Coprinus cinereus*, *Fusarium graminearum* (Cuomo, Guldener et al. 2007), *Fusarium oxysporum*, *Fusarium verticillioides*, *H. capsulatum*, *Magnaporthe grisea* (Dean, Talbot et al. 2005), *Medicago truncatula* (http://www.medicago.org/) *Rhizopus oryzae*,

S. Sclerotiorum, *Stagonospora nodorum* and *Toxoplasma gondii* were in various stages of genome sequencing and assembly.

3.2 Datasets for Supervised Training

Annotations of the *A. thaliana*, *C. elegans* and *D. melanogaster* genomes are obtained from TIGR Arabidopsis database (www.tigr.org), WormBase (www.wormbase.org) and FlyBase (www.flybase.net) respectively. A set of 1000 genes which are (i) validated by cDNA/EST mapping or confirmed by RT-PCR and (ii) not overlapping with test are selected from the annotation as training sets.

3.3 Datasets for Semi-Supervised Training

The set of *Saccharomyces cerevisiae* (1997) introns is downloaded from Ares lab (http://compbio.soe.ucsc.edu/yeast_introns.html). A total of 253 introns are reported in the Ares set. A subset of 185 introns and the corresponding BP motifs supported by experiment are used to determine the positional frequency models of donors, acceptors, and BP motif as well as the length distributions of intron and BP motif upstream and downstream spacers. The *S. cerevisiae* annotation (1997) is downloaded from GenBank to derive state transition probabilities and exon length distributions.

Table 3.1 shows the characteristics of *Hemiascomycetes* species that currently are under study.

Table 3.1 Species of *Hemiascomycetes* class. Significant variation in G+C composition observed within the same *genus* indicates evolutionary divergence and codon usage differences among these species.

species	**source**	**size (Mb)**	**link**	**G+C % of genomic sequence**
Ashbya gossypii	Ashbya Genome Database	9	http://agd.vital-it.ch/index.html	52
Candida albicans SC5314	Stanford	16	http://www.candidagenome.org	35
Candida glabrata CBS 138	Genolevures	12	http://cbi.labri.fr/Genolevures	39
Candida guilliermondii	Broad Institute	12	www.broad.mit.edu	44
Candida lusitaniae	Broad Institute	12	www.broad.mit.edu	45
Candida tropicalis	Broad Institute	15	www.broad.mit.edu	33
Debaryomyces hansenii	Genolevures	12	http://cbi.labri.fr/Genolevures	36
Kluyveromyces lactis	Genolevures	10	http://cbi.labri.fr/Genolevures	38
Saccharomyces cerevisiae	GenBank	12	http://www.ncbi.nlm.nih.gov/Genbank	38
Yarrowia lipolytica	Genolevures	20	http://cbi.labri.fr/Genolevures	49

3.4 Test Sets Preparation

To evaluate algorithm performance two types of test sets are derived. The first, Type I, include validated genes representing a set of genes with complete structure. The second, Type II, includes Type I and, in addition, contains gene structures partially supported by ESTs. Table 3.2 shows the size and structure for both Type I (a) and Type II (b) test sets. For all species with the exception of *C. reinhardtii* and *S. pombe* the test sets are derived by using alignments obtained from EST-to-genome mapping. For *C. reinhardtii* and *S. pombe* expert curated genes were used.

Table 3.2 The size and the structure of the test sets used for evaluation of algorithm performance.

a) test set Type I.

species	**genes**	**introns per transcript**
A. gambiae	144	2.4
A. thaliana	1,026	4.9
C. elegans	183	7.1
C. cinereus	167	3.4
C. immitis	432	2.3
C. intestinalis	314	7.0
C. reinhardtii	43	8.7
D. melanogaster	361	2.7
F. verticillioides	327	2.0
M. grisea	169	2.0
T. gondii	65	4.1
S. pombe	1,277	3.1

b) test set Type II.

species	**genes**	**introns per transcript**
A. nidulans	1,075	2.6
A. niger	955	2.8
A. terreus	729	2.8
B. cinerea	787	2.7
C. neoformans	2,425	3.8
F. graminearum	919	2.6
F. oxysporum	461	2.5
N. crassa	276	2.5
R. oryzae	2,169	3.3
S. nodorum	413	2.7
S. sclerotiorum	587	2.9

Note: neither set contains single exon genes.

3.4.1. Test sets derived by mapping

Mapping pipeline (Shmeleva N. et al., mid-size eukaryotes (unpublished research) and Kislyuk et al., fungal genomes (unpublished research)) provids candidate transcripts for the test set derivations. BLAT (Kent 2002) primary component of mapping procedure is used to align native (genome specific) ESTs to the genomic sequence. Then, high quality alignments with alignment identity and coverage better than 90% are chosen and clustered into transcripts that share common introns. These transcripts, the result of the pipeline, are further analyzed by a modified GeneMark.hmm program (Lukashin and Borodovsky 1998) with models described in (Besemer and Borodovsky 1999) for a presence of a non-interrupted protein coding region.

The resulting transcripts that satisfy the following conditions are included in the test set: (i) a gene should start with ATG and should contain canonical donor/acceptor sites (test set Type I); (ii) intron/exon structure should be supported by EST/cDNA alignment; (iii) no alternative isoform(s) should be present in the set; and (iv) there must be an in-frame stop codon that is positioned upstream of gene start or the length of upstream region must be greater than 150nt which reflects minimum error rate of heuristic model.

Sequences containing multiple genes per record (ideally, an accurately annotated region of a chromosome) are preferable for the accuracy assessment (Pavy 1999). Adjacent genes along with the intergenic region between them are accepted into the Test Set If this intergenic region has shown no similarity to the databases of ESTs and cDNAs sequences. Still, there is no guarantee that these precautions would result in intergenic

region that does not contain any coding region. Multiple genes per records are obtained for *A. thaliana* and *D. melanogaster.*

This procedure does not produce many regions with multiple genes adjacent to each other. Therefore, test sets for these species contain mostly one validated gene per record. Initiation and termination sites in the Type I test sets were determined computationally and therefore the accuracy values for these sites are within the error rate of the technique that was used to derive the set (estimated 1-3%) and should be considered cautiously.

3.4.2. Artificial Chromosomes

To assess the frequency of gene splitting and gene merging, an artificial *S. pombe* chromosome is constructed. *S. pombe* annotated proteins are blasted against the SwissProt database (Watanabe and Harayama 2001) with an e-value better than e-07. If the best hit of the alignments contains more than 95% of the length of both proteins, then this candidate gene is selected. Next, the genes are connected by random sequences of a particular length which reflect the non-coding region of the *S. pombe* genome (31% GC content). The "intergenic" sequences of a given "artificial chromosome" are chosen to have one and the same length *L*; thus, instances of the *S. pombe* artificial chromosome for several *L* values ranging from 50 to 6,000 nucleotides are created. All genes are placed in the direct DNA strand.

3.4.3. Test Sets for Genomes with a Small Number of Introns

The limited availability of ESTs and the strict requirements in test deriving procedure present challenge in acquiring reasonable size of multiple exon gene containing test sets for species where these genes are present in small numbers. Therefore, annotated multiple exon genes of four completely sequenced genomes (Dujon, Sherman et al. 2004) that satisfy the following criteria are included in the test set: (i) both the exon and intron lengths should not be less than 30nt; (ii) the record should contain complete gene structure; (iii) the gene should possess canonical signal sites; (iv) no in-frame stop codons should be present in the coding region; and (v) the spliced gene should not have alternative isophorm. Anotation of *Ashbya gossypii* (Dietrich, Voegeli et al. 2004) for contains regions labeled "unsure". For this genome an additional rule of was applied in test set selection to avoid genes that overlap with these regions. The test sets of *Candida albicans*, *A. gossypii*, *Debaryomyces hansenii*, *Kluyveromyces lactis* and *Yarrowia lipolytica* contain 240, 149, 138, 72 and 451 genes, respectively.

3.4.4. Externally Derived Test Sets

Genomic annotation has been previously used by Korf (Korf 2004) to derive test sets for *A. thaliana*, *C. elegans* and *D. melanogaster*. The author used a set of rules to eliminate possible errors in annotation, e.g. overlapping exons, exons out of bounds, mislabeled strand, in-frame stop codons, introns less than 30 bp. Moreover, each gene was confirmed by "an end-to-end, gap-free alignment between the *in silico* predicted transcript and a full-length cDNA" (Korf 2004). The test sets were downloaded from www.biomedcentral.com/content/ supplementary/1471-2105-5-59-S1.gz.

3.5 GeneMark.hmm E-3.0

Originally the GeneMark.hmm was developed for prokaryotic genomes (Lukashin and Borodovsky 1998) and then extended for eukaryotes (Lukashin A. and Borodovsky M., unpublished). Over the years different versions of GeneMark.hmm were developed. The program was successfully used in eukaryotic genome annotations (Initiative 2000, Yu, Hu et al. 2002). Statistical models for the gene finding algorithm described in this work are utilized by the latest version GeneMark.hmm E-3.0 (E- eukaryotic, 3.0-version) (Lomsadze, Ter-Hovhannisyan et al. 2005). The basis of GeneMark.hmm is a Hidden Markov Model (HMM) with duration or a hidden semi-Markov model (HSMM) (Rabiner 1989).

Formally, the HSMM for GeneMark.hmm E 3.0 is defined by the set of elements described below.

1. The state space $S = \{S_1, S_2, \ldots, S_N\}$. The state space is characterized by the finite number of $N = 56$ hidden states. The hidden states are initial, internal, terminal and single exons, introns, intergenic regions, and the boundaries between the states are the initiation, donor, acceptor, and termination sites. Figure 3.1 illustrates the HSMM diagram employed in the eukaryotic GeneMark.hmm E-3.0. Hidden states are shown for direct DNA strand only. A mirror symmetrical illustration of this diagram would present the hidden states generating a sequence of the complementary DNA strand. These hidden states generate sequence on both direct and complement DNA strands. Boundary states can be described by a variable length window and the splice sites reflect intron phase dependence.

Frequently for calculation purposes, two additional *Begin* and *End* states are introduced (see Viterbi algorithm).

2. The alphabet M = {A,T,G,C}. The alphabet consists of symbols emitted by a particular state.

3. Emission probability distribution $e_j(k)$. Each state emits a particular symbol with certain probability (hence emission probability) defined as follows

$$e_j(k) = P(X_i = k \mid S_i = j) \;\; k \in M \qquad \text{for } 1 \le k \le M,\ 1 \le j \le N \tag{3.1}$$

Equation (3.1) shows the emission probability for a zero order Markov model. For Markov models of higher orders, for example n, equation (3.1) becomes

$$e_j(k) = P(X_j = k \mid X_{i-1}, X_{i-2} \ldots X_{i-n} S_i = j) \;\; k \in M \tag{3.2}$$

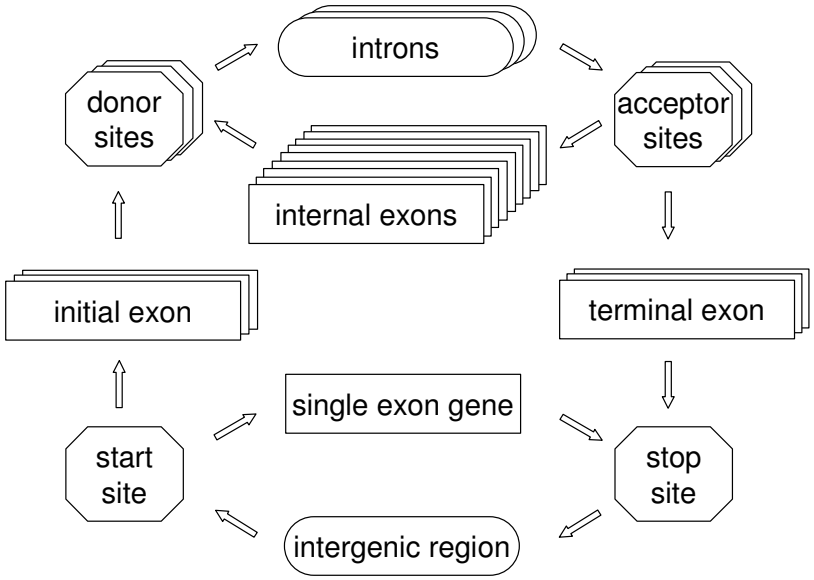


Figure 3.1 HSMM diagram used in GeneMark.hmm E-3.0. The states describing only direct DNA strand are shown. The reverse strand which is represented by reversed direction of arrows and mirror symmetrical image is omitted.

For each state the emission probability distribution is fixed regardless of which state or which point it arrived from. Sequences emitted by the non-coding (intron and intergenic region) and coding states are modeled by homogeneous and tree-periodic inhomogeneous Markov chains respectively (Borodovsky 1986, Borodovsky and McIninch 1993, Burge and Karlin 1997). Parameters of the intron and intergenic regions may be considered as separate states and can be described by different Markov models allowing asymmetry in introns. Depending on the size of the available training set, higher orders of Markov chain (normally up to five) can be implemented.

4. The state transition probability distribution $A = \{a_{S_j \to S_i}\}$, $a_{S_j \to S_i} = P(q_t = S_i \mid q_{t-1} = S_j)$, for $1 \le i, j \le N$ and q_t defining the current state. The difference between HSMM and conventional HMM is that the duration density is explicitly defined. State durations or length distributions ($P(d)$) are state-specific *e.g.* initial, internal, and terminal exons are described by the corresponding state's length distributions. The site states emit site type and intron phase dependent nucleotide sequences of fixed length modeled by positional (inhomogeneous) Markov chains. GeneMark-ES has the advantage of utilizing length distributions described by analytical functions, e.g. a geometric distribution, as well as by experimental observations, e.g. lengths of exons from the training.

5. The initial state distribution $\pi = P(q_1 = S_j)$, for $1 \le j \le N$ is limited to non-coding states which are characterized by their length distribution.

3.6 Finding the Best Gene Structure

GeneMark-ES employs dynamic programming, the Viterbi algorithm (Viterbi 1967), to determine the most likely sequence of underlying hidden states from observed nucleotide sequences. The Viterbi algorithm utilizes the memoryless property of Markov chains according to which probabilities of events do not depend on past *n* events (where *n* defines the order of Markov model). For a nucleotide sequence $x=(x_1,x_2\ldots x_L)$ of length L, where $x_1,x_2\ldots x_L$ represent the subsequence of L bounded by the site states, the highest scoring sequence of the hidden states is deduced in a four step procedure described below

Step 1. Initialization

A technical step to facilitate further calculations

$$\forall\ S_j\ j \neq 0 : U_{S_j}(begin) = 0 \quad U_{S_0}(begin) = 1$$

Set

$$U_{S_0}(begin) = 1$$

$$U_{S_j}(begin) = 0 \quad \forall\ S_j,\ s.t.\ j \neq 0$$

Step 2. Forward parse

Two variables reflecting (i) the probability of the most probable path of the states and (ii) information about the previous state up to the current point are stored in the memory.

$$\forall\, 1 \leq j \leq L : U_{S_i}(j) = e_{S_i}(x_j) P(d_{x_j}) (\max_S (U_{S_k}(j-1) a_{S_k \to S_i})$$

Step 3. Termination

The termination condition is fulfilled when *End* state is reached. The joint probability of the optimal path π^* and the most probable transition to the end state are determined by the following formulas.

$$P(x,\pi^*(S_i) = \max_S (U_{S_k}(L)a_{S_k \to S_{end}})$$

$$\pi_L{}^*(S_i) = \arg\max_S (U_{S_k}(L)a_{S_k \to S_{end}})$$

Step 4. Backtracking

In this step the optimal sequence is recovered by tracing back through the *ptr*

$$\forall i = (L...1) : \pi^*_{i-1} = ptr_i(\pi^*_i)$$

The calculations are done in *log* space to eliminate the problem of multiplying small values. The summation substitutes the multiplication in the four steps described above.

3.7 Gene Finding Programs Used for Comparison

SNAP (Korf 2004), a generalized HMM gene finder adaptable for novel genomes, is downloaded from http://homepage.mac.com/iankorf/snap-2005-07-27.tar.gz. Bootstrap models are not available within the distribution package. The supervised models are used in comparison with GeneMark-E and GeneMark-ES. The results from original publication are used (Korf 2004) to compare with bootstrap models.

AUGUSTUS, (Stanke and Waack 2003) another *ab initio* gene finder, is used to evaluate GeneMark-ES and GeneMark-ES-2 performance. Its latest version (AUGUSTUS 2.0) is downloaded from http://augustus.gobics.de/binaries/ and run locally using the default settings.

3.8 Heuristic Models

Heuristic approach described by Besmer et al. (Besemer and Borodovsky 1999) uses the compositional correlation between genomic G+C content and codon usage. This approach is utilized in the initialization of the self-training algorithm to describe the models of protein coding and non-coding regions.

3.9 Gibbs Sampler

The Gibbs sampler (Lawrence, Altschul et al. 1993, Thompson, Rouchka et al. 2003) is used in finding BP motif in the set of intron sequences. Local version of this motif-finding algorithm (version 3.02.003) is downloaded from http://bayesweb.wadsworth.org/ gibbs/gibbs.html. A site sampler mode that assumes single motif per sequence is used.

3.10 Accuracy Assessment

The accuracy of gene finders is evaluated using standard methods (Burset and Guigo 1996). Prediction sensitivity (Sn) and specificity (Sp) are defined as follows:

$$Sn = \frac{TP}{TP + FN} \tag{3.3}$$

$$Sp = \frac{TP}{TP + FP} \tag{3.4}$$

where,

TP (true positives) - number of correctly predicted features,

FN (false negative) - number of missed features,

FP (false positive) - number of extra predicted features.

Gene structure elements used in Sn and Sp calculation are (i) exons further categorized into initial, internal, terminal and all types; (ii) introns; and (iii) signal sites e.g. donors, acceptors, initiation and termination sites. These categories are not independent of each other, but each group assists in assessing the algorithm's overall performance. Although different measures describing both Sn and Sp have been defined (Burset and Guigo 1996) a simpler approach expressed as the average of these values is used here.

3.11 Sequence Logos

Sequence logos are widely used to visualize the consensus pattern of aligned sequences with a given alphabet, e.g., nucleotide or protein. This method was originally developed by in 1990 (Schneider and Stephens 1990). Logos display not only the consensus sequence but also some other important motif characteristics such as the relative frequency, the amount of information, and the order of significance of the residues at a particular position. Two programs that implement this approach are utilized by a software utility called "Pictogram" available at www.genes.mit.edu/pictogram.html and by a local version scripted by William Hayes. Here sequence logos are used to display the signal site signals including splice sites and the branch point site.

3.12 Other Scripts and Libraries Used in Self-Training

Several scripts earlier developed in Dr. Mark Borodovky's lab at the Georgia Institute of Technology are integrated into the model derivation step of the algorithm. Perl scripts are adopted and modified from Dr. John Besemer for data parsing and estimation of positional frequency models. Perl and C/C++ libraries written by

Aleksandre Lomdadze are used for length distribution smoothing and *a posteriori* probability estimation. Latest version of GeneMark.hmm, GeneMark.hmm E-3.0, used in this work is written by Aleksandre Lomsadze (Lomsadze, Ter-Hovhannisyan et al. 2005).

3.13 Server Description

All the computations were performed on high performance supercomputers. Initially, the GeneMark-ES was run on an IBM RS6000 8-CPU PowerPC 500 MHz server and a IBM P-series 2-CPU PowerPC 1.6 GHz server. These eight-way 64-bit SMP servers provided ECC L2 cache and Dynamic Processor Deallocation. On a genome with a size of 100 Mb, depending on the average server load GeneMark-ES employing multiprocessor mode runs for nearly 14 hours.

Later, with the installment of a new machine, the "topaz", GeneMark-ES-2 and GeneMark-LE were run on HPC Turnkey Beowulf-Class Supercomputers. With the Quad-Core Xenon architecture, this computer has 32x2-socket, 8-core compute nodes (Intel Xeon 2.33 GHz) with 8 GB RAM per node totaling 256 GB RAM and 4TB main storage.

The running speed for this computer is three-fold faster than for the IBM machine. The running time on topaz can be increased twenty-fold for the processes that can be parallelized by utilizing Portable Batch System (PBS) which utilizes all available nodes; process parallelization, however, is possible only for GeneMark.hmm E-3.0 run on the input sequences and currently is not feasible for the parameter re-estimation steps.

CHAPTER 4

SELF-TRAINING ALGORITHM GeneMark-ES FOR EUKARYOTIC GENE FINDING

4.1 Introduction

While the number of eukaryotic genome projects is growing at a fast pace the annotation process of a novel genome is lagging behind. The bottleneck is due to the fact that the majority of the gene prediction programs require a training set of genes verified by experiment in order to determine the model parameters. Frequently, genomic sequence but not the training set is readily available and awaiting annotation.

This chapter describes a self-training algorithm for eukaryotic gene finding. This method does not require a predefined training set in order to obtain model parameters for the predictor and eliminates the time and manual work invested into the compiling of the training sets necessary for traditional *ab initio* gene finding approaches. The method developed here solves the problem of the estimation of model parameters for a eukaryotic gene finder from unclassified (into coding and non-coding) nucleotide sequences. Particularly, this method can be applied to the novel genomes with insufficient amount of training data for model parameter estimation. In the theory of HMMs similar problems are commonly solved by an expectation maximization (EM) algorithm, the Baum–Welch algorithm (Dempster, Laird et al. 1977, Baldi 2000) which maximizes the likelihood function given the observed data. Convergence of the Baum–Welch algorithm to the global maxima is not guaranteed. Another alternative for the HMM parameter estimation

from data with missing features is the Viterbi algorithm (Viterbi 1967). Computationally, the Viterbi algorithm is less intensive than the Baum-Welch algorithm which requires significant computational resources for state duration parameter estimation. Viterbi allows analytical estimation of state durations from the predicted sequence parse.

Unsupervised training methods for prokaryotes that employ Baum–Welch or Viterbi algorithm have been developed (Delcher, Harmon et al. 1999, Baldi 2000, Besemer, Lomsadze et al. 2001). GeneMarkS (Besemer, Lomsadze et al. 2001), an iterative unsupervised training procedure that starts the iterations with heuristically described models (Besemer and Borodovsky 1999), derives statistical models for prokaryotic GeneMark.hmm. Similarity searches were integrated into the self-training algorithms such as ORPHEUS (Frishman, Mironov et al. 1998) and EasyGene (Larsen and Krogh 2003). Self-training has been used to identify atypical genes acquired by the genome in the course of evolution by horizontal transfer (Hayes and Borodovsky 1998).

Until recently there was an opinion that generation of accurate gene modes via unsupervised training was not possible for eukaryotic species with more complex gene structure. Indeed, the increased number of parameters and the noise introduced by the abundance of non-coding DNA create major challenges for eukaryotic gene finding.

The proposed method, GeneMark-ES, is first of a kind unsupervised training procedure for eukaryotic genomes. The gene prediction and the model parameter estimation is implemented in iterative fashion. As iterations progress, the algorithm recurrently parses the input sequence into coding and non-coding regions (the expectation step) and estimates new model parameters (the maximization step) from this parse. A

number of constrains are imposed on training process to filter out possible noise and to minimize the chance of converging to a biologically irrelevant point in the parameter space. The HSMM with reduced architecture is employed at the beginning of iterations and then gradually increases its complexity in further iterations. The algorithm was tested on well-studied genomes of *A. thaliana*, *C. elegans* and *D. melanogaster*. The observed gene prediction accuracy is comparable or better than the accuracy of a conventional method. The algorithm was applied to eukaryotic species with genomes incomplete at a the time the research was conducted.

The species presented in this chapter vary in their genome size, ranging from 80 Mb (*T. gondii*) to 270 Mb (*A. gambiae*), and G+C content, observed lowest at 35% (*C. elegans*) and highest at 63% (*C. reinhardtii*).

The algorithm produces models with acceptable prediction accuracy from the input sequence of 10 Mb (Section 4.3.4). GeneMark-ES predicted proteins were compared to the set of annotated proteins to identify genes missed in annotation. The algorithm is currently used by several research groups and sequencing centers. Particularly, it is part of the nematode annotation pipeline at wormbase (www.worbase.org), used for eukaryotic genome annotation at DOE Joint Genome Institute and Munich Information Center for Protein Sequences (MIPS).

4.2 Methods

This section describes the GeneMark-ES procedure. A block diagram of the automated eukaryotic self-training procedure GeneMark-ES is shown in Figure 4.1. The key steps of the algorithm are (i) initialization of HSMM parameters; (ii) parsing of the

input sequence into "coding" and "non-coding" regions by GeneMark.hmm E-3.0 with the model obtained in previous step which provides information for further refinement of the sets of labeled sequences and for updating the estimates of the model parameters to be used in the next iteration (details described in Sections 4.2.2 and 4.2.3); (iii) HSMM parameter re-estimation using the subset of the parse from (ii). Steps (ii) and (iii) are repeated until the algorithm convergence. Gene model obtained at the convergence step is used to produce the final predictions.

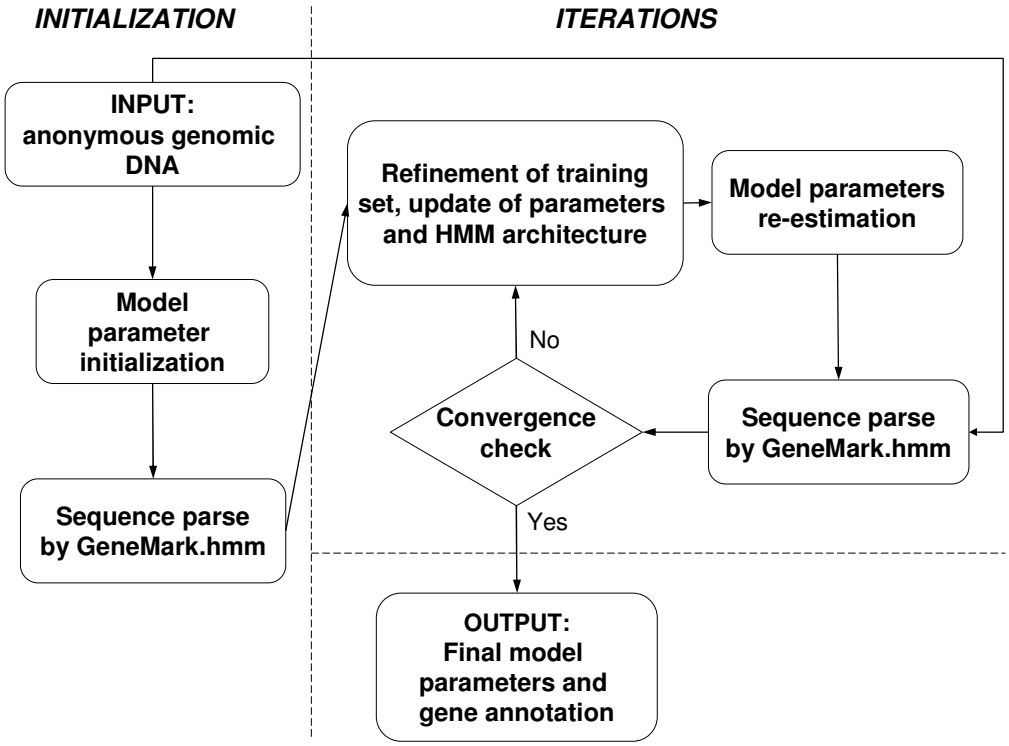


Figure 4.1 The step-wise diagram of GeneMark-ES, a self-training algorithm for eukaryotic species.

4.2.1. Initialization of Model Parameters

To initialize the model parameters for the protein coding state three approaches are considered:

1. Heuristic initialization

Three-periodic second order Markov chain models with parameters inferred from codon frequencies determined as functions of genome G+C content (Besemer and Borodovsky 1999). The non-coding states are described by a zero order Markov model which reflects the G+C content of the input sequence.

2. ORF initialization

The fifth-order inhomogeneous Markov chain with parameters derived from the sets of non-overlapping ORFs longer than 1000 nt are obtained form the input sequence. The non-coding states are described as for heuristic initialization

3. G+C fixed value initialization

A model based on a homogeneous zero order Markov model with GC content elevated by 8% in comparison with the genome G+C content; for coding and non-coding states described by a homogeneous Markov model with G+C content decreased by 4%. For example, if the average G+C for a particular sequence is 50%, then the probability of nucleotide 'G' being emitted by the coding state is $P(G_{f1}) = P(G_{f2}) = P(G_{f3}) = 0.29$, and $P(G) = 0.21$ being emitted by the non-coding state where f is the codon position or the frame. The main purpose for this initialization method is to test the algorithm robustness to a weak starting model. The simple justification for the parameterization is that the "G" and "C" nucleotides are on average more frequently observed in protein coding regions than in non-coding DNA.

The model at advantage is the heuristic initialization (1) which on average converges the fastest. The results of GeneMark-ES reported here are based on the heuristic parameter initialization of protein coding and non-coding states. A detailed discussion and comparison of different initialization models described above are presented in Section 4.3.5.

The following assumptions are made for the initialization procedure.

1. Due to the lack of knowledge, transition probabilities between states are assumed to be equally probable.

2. The state durations are defined by uniform distributions and limited by the state's minimum and maximum length settings (see Section 4.2.6).

3. Only canonical signal sites are considered (e.g. donor can be represented only by "GT" dinucleotides).

4. The site state models are presented by minimum (universal) signal sites with no information describing the upstream and downstream sequences of a given motif, *e.g.* donor (acceptor) site states emit two canonic GT (AG) dinucleotides, initiation (termination) site states emit canonic triplet sequence ATG (TGA, TAG, TAA).

5. Exons flanked by introns in phases 0, 1, and 2 are assumed to be accruing with equal probability.

4.2.2. Training Set Refinement

At a given step of iterations the training set is defined by the algorithm predictions with the model derived from the preceding step. For instance, after the algorithm is run for the first time with the heuristic model, the resulting predictions are used to create a

new gene model and the next round of iterations starts with the next parse of the input sequence. The predictions however contain noise and certain amount of false predictions. Filtering techniques are applied to the resulted parse of the current iteration to decrease the amount of mislabeled genomic sequences in the training set in the course of iterations.

The rate of false positive predictions is higher for shorter protein coding regions than for long coding sequences. Genes that contain a short coding sequence (CDS) are not considered in the parameter re-estimation step; if a predicted gene produces a CDS shorter than the length of N nucleotides (defined by the length of available data for training; minimum 300 nt, maximum 800 nt) the elements of this gene are removed from the training set. The 5' UTR and 3' UTR are removed from the training set and not considered in non-coding state parameter estimation.

Frequently, the input DNA sequence is available in the form of contigs or scaffolds, which are the result of an intermediate step of the eukaryotic chromosome assembly process. The 5' and 3' ends of the contigs/scaffolds are detached from the rest of the genome sequences making it difficult to determine with high confidence whether the genes predicted at both ends of the contig are complete or incomplete. To avoid such cases, the first and the last predicted genes, are not included into the training set.

Non-nucleotide characters, the result of the sequence assembly process, may introduce noise and considerably influence the parameter estimation process especially for protein coding region. If a subsequence corresponding to the particular state contains such characters, it is not considered in parameter estimation step.

It is important to ensure that the subset of the training set selected for parameter estimation of a particular type does not overlap with the subset of different type.

Consider, for example, an intron of length 20 nt, a positional frequency models with width 4 and 18 nt for donors and acceptors, respectively, are emitted by non-coding region. In this case, both splice models will have an overlap of two nucleotides and inclusion of these splice site into training set will introduce an error. Therefore these particular splice sites are removed from the training of donor and acceptor site models.

4.2.3. Parameter Constraints and Space Restriction

Mitrophanov *et. al.* have shown that the variations in estimation of emission probabilities have greater impact on HMM performance than variations in estimation of transition probabilities (Mitrophanov 2005). The signal sites contribute significantly to gene prediction accuracy as well (Figure A1 in Appendix). For this reason, the self-training algorithm starts the iterations with state emission probabilities and the site state model parameter re-estimations only. The splice site models are intron phase independent in the initial iterations. As iterations progress and these parameters gain discrimination power, the update of remaining parameters, such as the state durations for protein coding and non-coding regions, intron phase dependency, is allowed. These parameters include the state durations, the signal site and exon dependence on intron phase. The number of iterations when all other states are "freed" is set to three.

At a given step of iterations the site states emit a fixed number of nucleotides. The length (window) is determined based on the amount of information that a given position carries. In iterations these lengths are subject to a change. Acceptor site signal, for example, carries-position specific information within intronic region which gradually increases towards the dinucletides "AG". The Kullback-Liebler (KL) distance (Kullback

1951) between the acceptor motif and the background (introns) is used to evaluate whether it is expedient to include a particular position in the acceptor site model. KL is an indication of information divergence and is calculated by using the formula

$$D_{KL}(P \parallel Q) = \sum_{j}^{L} \sum_{i=A,T,G,C} P(i,j) \log \frac{P(i,j)}{Q(i)}$$

where

P(i,j)- probability estimate of nucleotide *i* in position *j* of the motif, and

Q- probability of a nucleotide *i* in non-site region (*e.g.* intron).

Note: to express the KL distance in bits logarithm of base 2 is used.

The position *x* for which the KL value is less than the mean KL of window N is set as the site state's border. Number of nucleotides between this position and a particular site is used as the length of the site state for the next iteration.

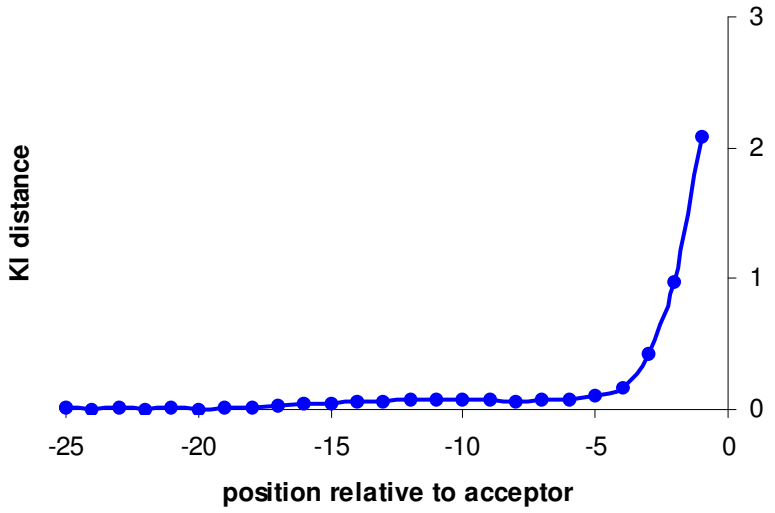


Figure 4.2 Zero order Kullback-Liebler distance between *A. thaliana* acceptor site models and intron model as determined from the training set at the algorithm convergence.

The change of values of the KL distance upstream of the acceptor site is shown in Figure 4.2. The site state length falls to the same window length for other species (Figure A2 of Appendix).

The maximum variation in the window length observed is ±1 nt does not cause significant differences in the final output of the procedure reflected in the exon-intron structure of the predictions. Therefore, to avoid additional steps and to decrease the computational time, the window size of the site states are fixed throughout iterations in the distribution package of GeneMark-ES.

Prediction error rate and subsequently error in parameter estimation is higher at the beginning of iterations. In addition to filters described above the minimum state duration for coding and non-coding is set to 20 nucleotides in the three initial steps of the algorithm.

4.2.4. State Durations and Transition Probabilities

As the iterations begin, the transition probabilities and state durations are assumed to be uniformly distributed for all states and are not updated in the initial two iterations, after which the state duration parameters are "freed". State-specific length distributions are explicitly defined for exons and introns using the lengths observed in the training set (see Section 4.2.6 for more details). Length distributions for introns and initial, internal and terminal exons are derived from corresponding data in the training set. To enrich the set of lengths for calculation of single exon genes as well as to minimize weight of

randomly occurring ORFs falsely predicted by the algorithm single exon length distribution is derived from the lengths of all CDS of predicted genes at a given iteration.

For each of 10,000 points describing exon length, for example, there is a need of about 100 data points which is unfeasible. In addition, for a given state the set of observed lengths used to derive state duration contains noise. Frequently, due to incomplete sequencing project and/or filtering methods, the lengths observed in the training set, after each iteration, may not be representative of the whole complement of length distributions. To overcome difficulties in estimation of the density function, smoothing and approximation methods such as nearest neighborhood are usually used, e.g. see (Silverman 1986). For coding states this is implemented in two steps: (i) to remove the three-periodicity, the length frequencies are averaged over bin of three nucleotides, and (ii) a smoothing algorithm is applied. For introns and single exon genes step (i) is irrelevant and only step (ii) is applied. Figure 4.3 shows the result of the smoothing applied to the lengths of exons observed in *A. thaliana* at the final step of GeneMark-ES.

4.2.5. Convergence

The difference between the iterations can be characterized by the nucleotide identification sensitivity and specificity values (Sn and Sp) with 'gene annotations' defined by the prediction parse from the previous iteration and 'gene predictions' which in their turn are defined by the sequence parse of the current step.

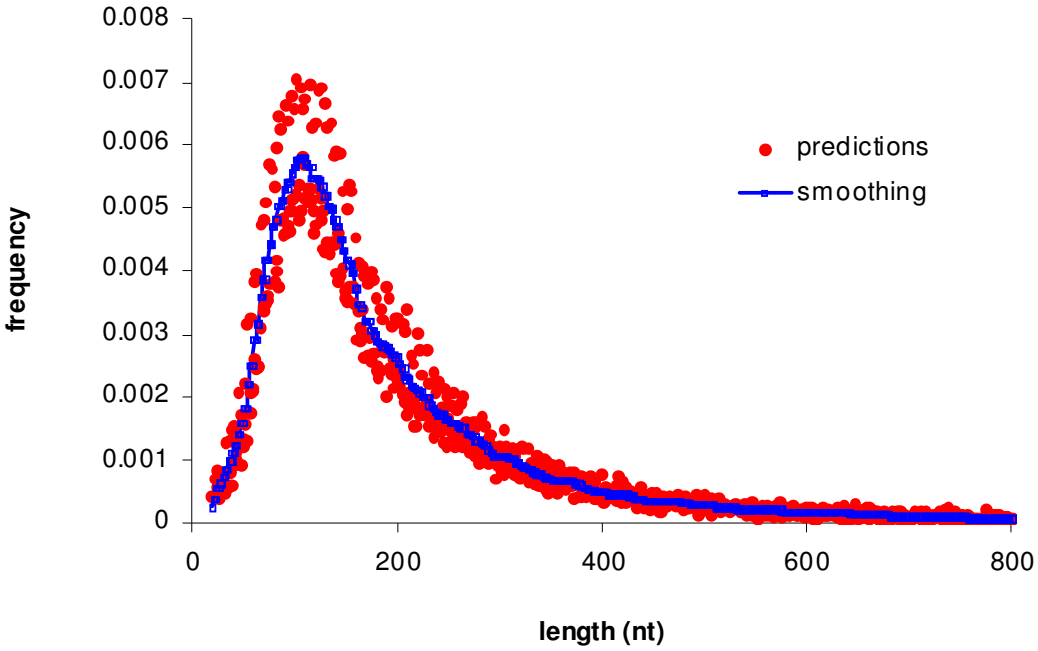


Figure 4.3 *A. thaliana* exon length distribution as obtained from GeneMark-ES predictions at the convergence (red dots) and from the result of the smoothing algorithm applied to the observed data (blue line).

The condition for algorithm termination is defined as average the Sn and Sp values in terms of nucleotide prediction accuracy. The automatic training procedure is considered to reach a convergence as the average values of Sn and Sp exceeds 97% in the category of splice site identification. The final output parse defines the predicted exon–intron structures while the values of parameters of the models derived from the final training parse are considered to be the final parameter estimates.

4.2.6. Thresholds and Settings

Several thresholds imposing restrictions on parameter space are introduced. The main purpose of defining the thresholds is to reduce the rate of the false positive predictions and to decrease complexity of computations.

As mentioned above, in the initial three steps of the algorithm minimum state durations were set to 20 nt to minimize the rate of false positive predictions in early iterations. Minimum threshold for single exon genes is set to 300 nt. The maximum durations are bounded by the length of the longest ORF that is observed in the given genome and 10,000 nt for sequences emitted from coding and non-coding states respectively. The 10,000 nt limit is chosen to control the algorithm's run time and can be increased on high performance computers or for those genomes that contain a high ratio of long intergenic regions. Note that the latter is often unknown for novel genomes but can be inferred based on the available information e.g. genome size, data from the close relatives. As the iterations progress, the state's minimum thresholds are switched to 3 nt for initial and terminal exons, 20 nt for internal exon lengths and 300 nt for a single exon genes.

A refinement procedure (Section 4.2.2) and frequently a relatively small size of input data may lead to a training set which represents an incomplete sample from the whole set. The low frequency of a particular parameter, for example, can be a consequence of an inadequate data supply which will cause the given parameter underestimation. This problem is often solved by use the of pseudo-counts. In this case, all parameters are required to be non-negative and non-zero (except when biologically

relevant e.g. emission probability of the in-frame "TTT" codon as a translation initiation start).

Set of fixed parameter settings based on *a priori* knowledge is introduced. These parameters are intron phase probability (initially set to uniform distribution), probability of the non-nucleotide character ("N" or "n" closing gaps between contigs in the sequence assembly process) being emitted by coding and non-coding states. In the later version of the algorithm intron phase probability is dynamically estimated in iterations (library functions added by Lomsadze, A). Given the inhomogeneous nature of the coding state the length of site states in the protein coding region is limited to 3nt in all iterations.

4.2.7. Pre-processing of the Input Sequence

Raw sequence data often contains features undesirable for model training. Repetitive sequences and low complexity regions for example cause problems in the DNA sequence assembly process, and by their nature these structures introduce bias in model parameter estimation. Gaps in the assembled DNA sequence are filled with non nucleotide characters (the most common being 'N') which can cause the algorithm to follow the path of false predictions such as merged, split or incomplete genes. Short contigs frequently correspond to a low coverage region and may contain a higher error rate.

Although a topic of separate research, the data pre-processing was reduced to relatively simple steps. A total of 1 Mb of sequence was removed from 5' and 3' of chromosomal DNA (if chromosomal assembly was available) to eliminate sequences which may possibly belong to telomeric regions rich in repetitive elements. In addition,

sequences that belong to sex chromosomes were not considered for the training process of *C. elegans* and *D. melanogaster*. In contrast, to autosomes the sex chromosomes are less stable and are poorly populated with genes, and the non-coding region contains a large number of repetitive elements and sequence duplication (Pimpinelli, Berloco et al. 1995, Charlesworth 1996, Orr and Kim 1998, Reinke, Smith et al. 2000, Carvalho, Dobo et al. 2001, Balakirev and Ayala 2003, Skaletsky, Kuroda-Kawaguchi et al. 2003). Recent studies show differences in gene content and expression pattern between *D. melanogaster* X chromosome and the autosomes (Vicoso and Charlesworth 2006). The non-recombining and degenerative Y chromosome in fruit fly has been shown to contain a low number of genes (Carvalho, Dobo et al. 2001) and genetic functionality (Charlesworth 1996, Orr and Kim 1998).

The input sequences were cleaned from long (greater than 1000nt) stretches of non nucleotide symbols. Contigs with length shorter than *L* nucleotides were not used in training. The *L* was chosen so that the size of the remaining set of sequences would at least be 10 Mb. Figure 4.4 shows the amount of the DNA sequence available for a particular cut-off length applied to the set of raw sequences of *T. gondii*. The cut-off length of 150 Kb, for example, provides set of sequences with size of 19 Mb.

One of the major problems with the input DNA sequences was observed for *Histoplasma capsulatum* which was in the initial stages of sequencing. The G+C graph for this genome shows inhomogeneous distribution with two maximums at 29% and 47% (Figure 4.5). More than 30% of the sequence data falls within the low G+C category. For this genome it was known by the representatives from the sequencing center (the Broad Institute) that the low end of the distribution corresponds to non-coding DNA.

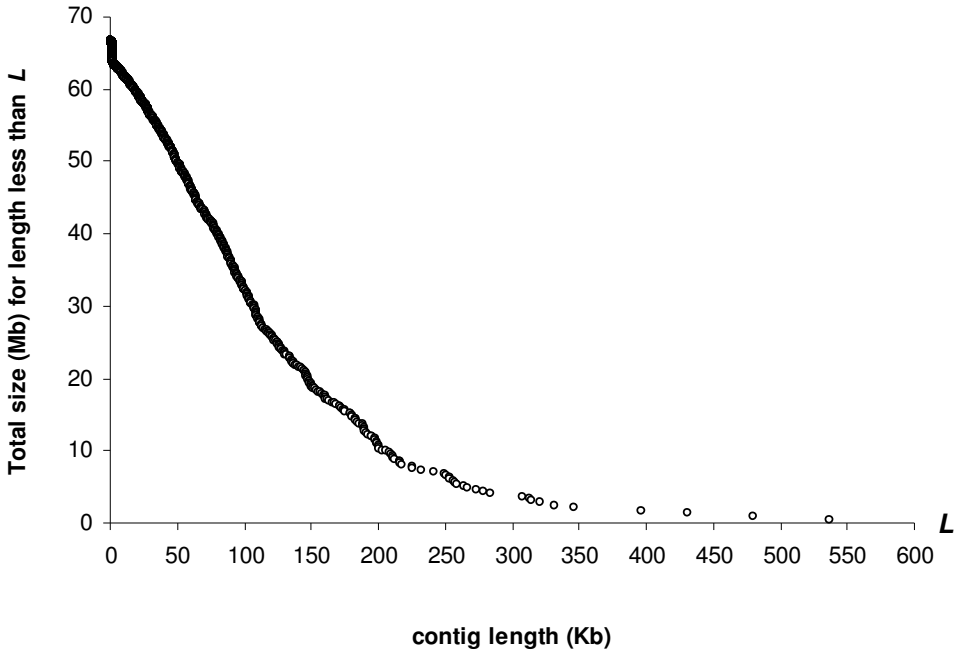


Figure 4.4 Total size of *T. gondii* sequences as a function of contig length. The set of sequences with total size of 19 Mb is obtained for contigs with lengths greater than 150 Kb.

This information came from the G+C distribution of the EST which did not contain sequences with low values of G+C as well as absence of EST alignments to this (low G+C) region (personal communication). Therefore, the genomic sequences were screened and low G+C regions where removed from and the remaining subset used for unsupervised training.

4.2.8. Comparison with Annotation and New Gene Identification

Final predictions of GeneMark-ES are used to compare with the publicly available sets of annotated proteins (*set A*) to find novel genes. First, the predicted genes are translated into protein sequences (*set B*). Then, a locally installed program, *blastp*

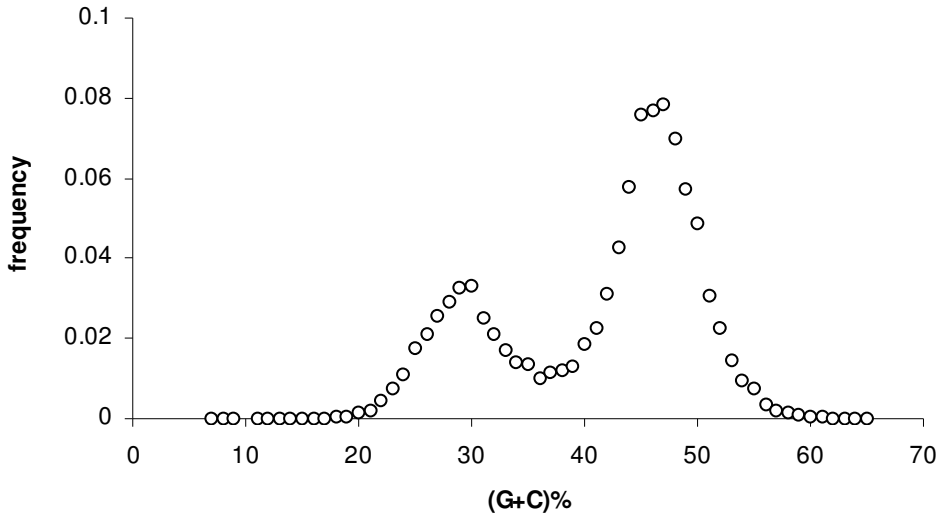


Figure 4.5 G+C histogram of genomic DNA determined for *H. capsulatum*. The histogram is calculated for 1 kb long non-overlapping fragments and shows two distinct peaks with maximums at 29% and 47%. Low G+C regions which correspond to non-coding DNA were removed from training.

(Altschul, Gish et al. 1990), is used to search for similarity between *set A* and *B*. The significance of the alignment defined by its e-value is required to be better than 10^{-5}. Next, the proteins from *set B* are selected into the *set C* so that the proteins in *set C* do not have a hit to any of the sequences in *set A*. *Set C* is searched against a non-redundant protein database (NR); this database is created by utilizing sequences from a number of databases including GenBank CDS translations, RefSeq (Pruitt, Tatusova et al. 2007), PDB (Berman, Westbrook et al. 2000) and SwissProt (Watanabe and Harayama 2001). In addition to e-value constraints, thresholds on the length of the alignment are imposed. The protein is labeled as novel if the alignment (i) satisfies the e-value threshold and (ii) contains more than 80% of the length of both the query (predicted protein) and its first (the best) hit from NR database. The novel proteins are then used to search for

conserved domains against the CDD (Marchler-Bauer, Anderson et al. 2003) to determine the functionally important regions in predicted proteins.

4.3 Results and Discussion

The result of GeneMark-ES when applied to genomic sequences of *A. gambiae*, *A. thaliana*, *C. elegans*, *C. intestinalis*, *C. reinhardtii*, *D. melanogaster* and *T. gondii* are presented and discussed in this section.

4.3.1. Algorithm Accuracy Evaluation

The algorithm performance is tested on a set of genes for which the exon intron structure is supported by EST/cDNA (see Chapter 3). For well-studied species such as *A. thaliana*, *C. elegans* and *D. melanogaster* model parameters are generated form the set of available genes to serve as a benchmark.

The Table 4.1 shows the accuracy results reflected in Sn, Sp, and their average values for both training methods, supervised and unsupervised. In a total of twelve categories, self-training produces better results than supervised training in seven cases for *A. thaliana* (in another three categories the same results are observed), eight cases for *C. elegans* and six cases for *D. melanogaster*. For these species the new method demonstrates splice site sensitivity and specificity values exceeding 92.8% and 87.0% respectively. Notably, for *C. elegans* the nucleotide sensitivity reaches 99.1%.

The results indicate that on average self-training performs better than or comparable to supervised training and thus can be successfully applied to novel genomes.

Table 4.1 Sensitivity and specificity (Sn/Sp) values and their averages for several categories of gene structure elements, characterizing the accuracy of gene prediction by GeneMark.hmm with models derived by unsupervised and supervised trainings. Bold font shows the larger value out of the two in corresponding category between supervised and unsupervised training methods.

		A. thaliana				*C. elegans*				*D. melanogaster*			
		supervised		unsupervised		supervised		unsupervised		supervised		unsupervised	
Internal exon	Sn	**91.2**	**89.9**	**91.2**	89.5	90.9	90.9	**94.0**	**92.7**	87.2	88.7	**91.3**	**91.0**
	Sp	**88.5**		87.8		90.8		**91.3**		90.2		**90.6**	
Donor	Sn	**94.0**	91.9	**94.0**	**92.2**	93.7	92.6	**96.2**	**93.5**	91.3	90.2	**92.9**	**90.3**
	Sp	89.8		**90.3**		**91.4**		90.8		**89.1**		87.7	
Acceptor	Sn	93.6	91.4	**94.0**	**92.1**	95.2	94.0	**97.3**	**94.5**	90.5	89.2	**93.2**	**90.4**
	Sp	89.2		**90.2**		**92.8**		91.6		**87.9**		87.5	
Initiation site	Sn	**80.1**	76.0	**80.1**	**78.3**	79.2	73.3	**85.8**	**77.4**	83.4	78.9	**84.5**	**79.1**
	Sp	71.9		**76.5**		67.4		**68.9**		**74.3**		73.7	
Termination site	Sn	**88.3**	83.5	87.5	**85.3**	94.0	**86.8**	**95.1**	85.2	89.5	**84.2**	**89.8**	83.6
	Sp	78.6		**83.1**		**79.6**		75.3		**78.8**		77.3	
Nucleotide	Sn	97.2	95.8	**97.7**	**96.3**	97.8	**96.7**	**99.1**	96.4	**98.1**	**95.6**	98.0	95.4
	Sp	94.3		**94.8**		**95.5**		93.6		**93.1**		92.8	

Note: The results are based on test set Type I as described in Chapter 3.

Ideally, when the well annotated whole set of genes is used in supervised training the difference between these two methods is expected to be minimal. The highly expressed genes, usually the core of the supervised training set, introduce bias in supervised training. This is not the case, however, for the unsupervised training.

As described in Section 4.2.7 the sex chromosomes were excluded from the training due to the high levels of noise. The algorithm exhibits marginal improvement in accuracy when training is employed without *D. melanogaster* X chromosome (Table 4.2).

Due to the lack of known genes in novel genomes, the supervised training is not employed for *A. gambiae*, *C. intestinalis*, *C. reinhardtii* and *T. gondii*. The performance of the unsupervised training is assessed on the test set only. Table 4.3 shows the prediction accuracy of GeneMark-ES for novel genomes.

Table 4.2 Accuracy of the self-training algorithm in terms of Sn and Sp and their average with two types of inputs of *D. melanogaster* sequences. Self-training shows marginal improvement when trained without sequences of the X chromosome. Bold font shows the larger value out of the two in corresponding category between two unsupervised training methods.

		D. melanogaster			
		excludes X chromosome		includes X chromosome	
Internal exon	Sn	**91.3**	**91.0**	**91.3**	90.5
	Sp	**90.6**		89.7	
Donor	Sn	**92.9**	**90.3**	92.8	90.0
	Sp	**87.7**		87.2	
Acceptor	Sn	**93.2**	**90.4**	93.0	90.0
	Sp	**87.5**		87.0	
Initiation site	Sn	**84.5**	79.1	83.9	78.7
	Sp	**73.7**		73.5	
Termination site	Sn	**89.8**	**83.6**	89.2	83.2
	Sp	**77.3**		77.2	
Nucleotide	Sn	**98.0**	**95.4**	97.9	**95.4**
	Sp	92.8		**92.9**	

Note: The results are based on test set Type I as described in Chapter 3.

The results for *C. reinhardtii* are among the highest in terms of Sn and Sp. *C. reinhardtii* is a single cell green algae found in soil and fresh water. What sets this species apart is its high 63% G+C genome content and consequently frequent G or C occurrence in the third position of the codon, creating a highly biased codon usage for this species. This helps to better discriminate between coding and non-coding sequences providing more accurate predictions in iterations.

The algorithm performance on *T. gondii*, where error rate in nucleotide identification reaches nearly 10% (Table 4.3), is related to two main issues. The first and most likely the main reason, is the greater than usual fraction of non-canonical splice sites estimated to be nearly 10% (Berriman, M., personal communication). Another

Table 4.3 Sensitivity and specificity (Sn/Sp) values and their average for several categories of gene structure elements characterizing accuracy of gene prediction by GeneMark.hmm with models derived by unsupervised training for novel genomes.

		A. gambiae		*C. intestinalis*		*C. reinhardtii*		*T. gondii*	
Internal exon	Sn	89.3	88.9	94.8	93.5	91.4	93.4	80.2	81.7
	Sp	88.4		92.1		95.4		83.1	
Donor	Sn	89.7	86.9	95.3	92.5	94.1	95.2	81.3	84.4
	Sp	84.1		89.7		96.3		87.5	
Acceptor	Sn	92.3	88.5	96.3	93.3	93.5	94.6	82.0	85.2
	Sp	84.7		90.3		95.7		88.3	
Initiation site	Sn	77.8	72.9	79.6	71.3	82.9	78.4	58.5	65.1
	Sp	67.9		63.0		73.9		71.7	
Termination site	Sn	86.1	78.9	85.4	75.9	92.7	87.7	66.2	73.7
	Sp	71.7		66.3		82.6		81.1	
Nucleotide	Sn	96.0	90.5	98.3	94.2	97.4	97.4	89.6	88.4
	Sp	85.0		90.0		97.4		87.1	

possible reason is that the shape of intron distribution which shows two peaks (Figure 4.6). The first maximum is highly localized at 38 nt, which may be side-effect of the erogenous predictions cause by non-canonical splice sites, while the second at ~450nt, being the longest among species described here, is widely spread. Mislabeling such long introns in turn greatly affects the nucleotide prediction accuracy.

Two peak intron distributions may occur in the course of unsupervised training as a side effect if the genomic sequence has low coverage and frame-shifts caused by the sequencing errors in protein-coding regions. To demonstrate this, artificial frame-shifts is introduced into genomic sequence of *A. thaliana* by randomly deleting a base in the genomic sequence with a frequency of 1/1000 nt. Then the self-training is applied to the modified genomic sequence. At the algorithm convergence, the length distribution of introns (Figure 4.7), but not that of exons were affected by creating an extra peak in the short intron range (32-45 nt) which coincides with that of *T. gondii*. In addition

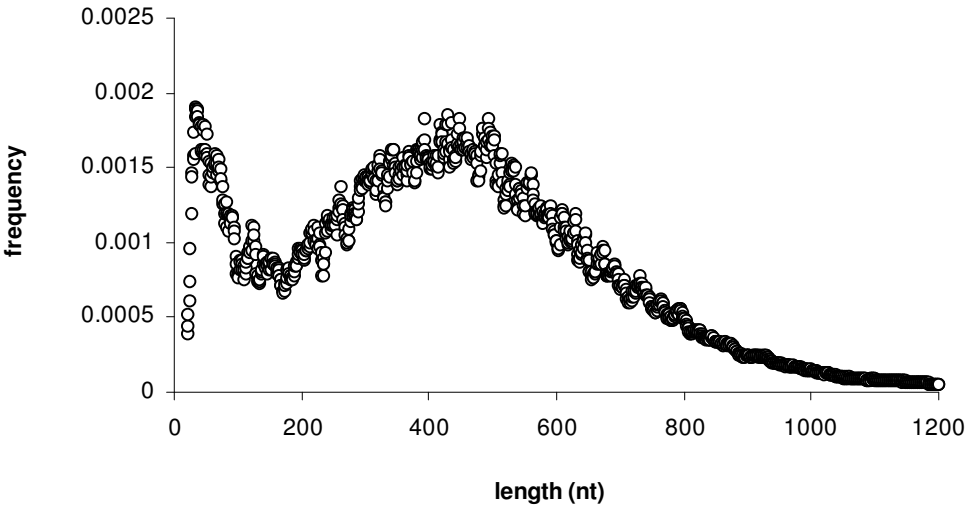


Figure 4.6 The intron length distribution of *T. gondii* exhibits two peaks. First, with maximum at 38nt is highly localized and the second with maximum at 450nt is less skewed.

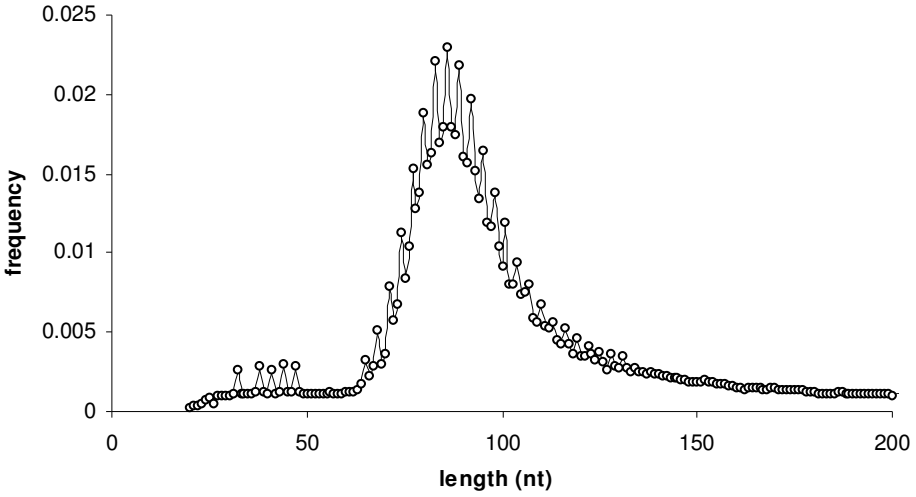


Figure 4.7 The intron length distribution obtained from the run of GeneMark-ES on modified sequences of *A. thaliana* shows a distinct noise/peak in the range of 32-45nt. 3k+2 periodicity observed in the range of 65-125nt suggests that the deletions in coding region cause frame shifts leading to artificial intron retention.

periodicity of the distribution favoring 3k+2 (k = 21, 22,…, 43) lengths is observed in the range of 65-125 nt. The intensity of the periodicity is proportional to the rate of mutations. The algorithm has finds most of the coding region correctly, but in the regions where the deletion was introduced, it is typically forced to predict a short intron.

4.3.2. Dynamics of Convergence in Iterations

To track the progress of the unsupervised training, GeneMark.hmm E-3.0 is run on the test set with the model that is produced at the current step of iterations to determine its Sn and Sp values. The results show that as expected the initial models producing 5% to 40% accuracy are the weakest. Figure 4.8 shows the prediction accuracy of gene elements as determined from the test sets of *A. thaliana*, *C. elegans* and *D. melanogaster*. These parameters are shown as functions of iteration index. For gene predictions produced by models defined at initialization, the Sn and Sp values are shown at the index value one. Similar results are observed with the application of GeneMark-ES to genomes of *A. gambiae*, *C. intestinalis*, *C. reinhardtii* and *T. gondii*.

As the algorithm follows the process of iterations, Sn and Sp values grow. The rate of the growth is high during initial iterations especially for the models derived at the first iteration where the site state models containing minimal universal signals are switched to the full window mode (excluding the phase dependence for the splice sites) and Markov chains described by the heuristic method are updated for the first time. The growth is even more dramatic for specificity values, averaging a gain of nearly 40% after the first iteration. This trend is important for the training process as the higher the Sp

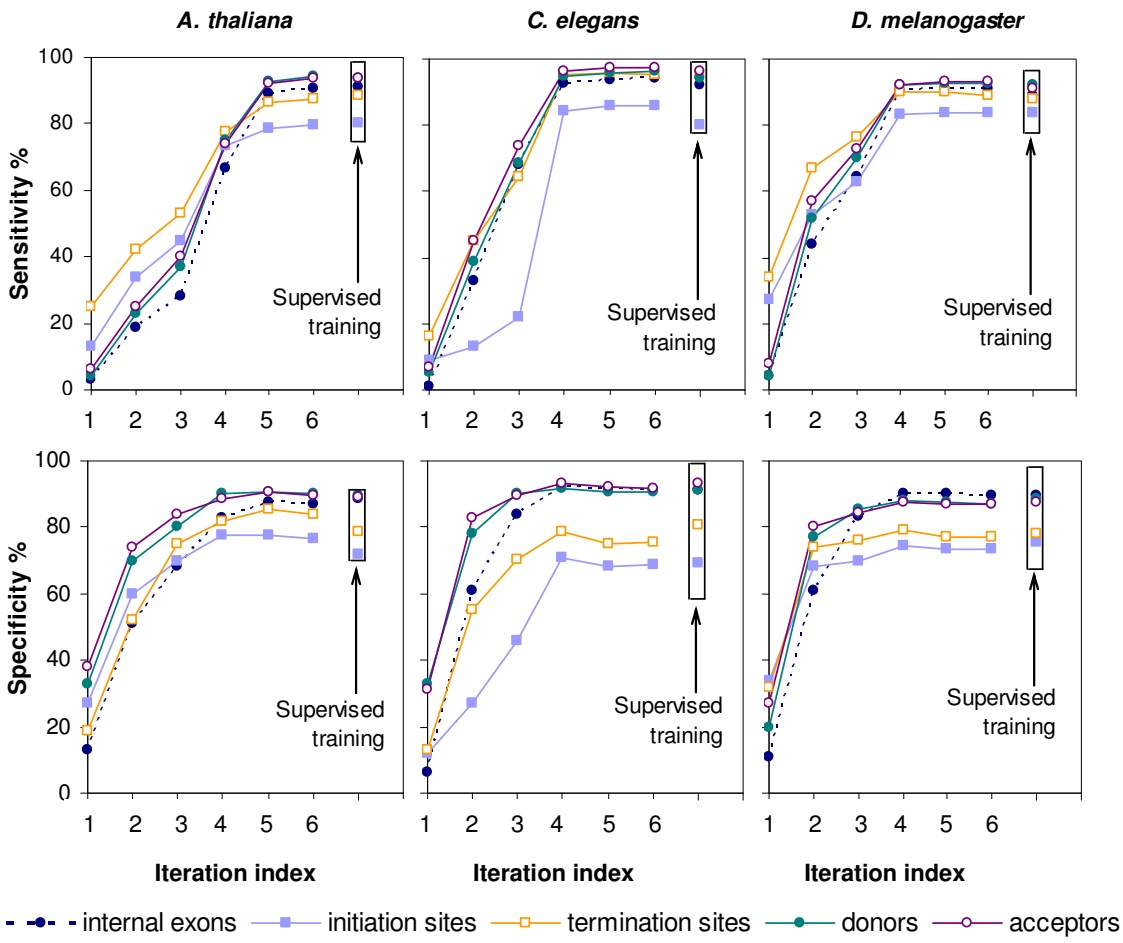


Figure 4.8 GeneMark-ES prediction accuracy (Sn/Sp) as a function of iteration index for three well-studied eukaryotic genomes. The Sn and Sp is determined at each step of iteration as tested on the test set Type I. Weak heuristic models provide with the initial parse of genomic sequence. Subsequent rounds of iterations and data refining enrich the training set with true positive predictions. At the convergence the accuracy results (Sn/Sp) are among the closest to biologically relevant point.

value, the fewer false predictions will be selected for the training set. For all species within three to four iterations, the algorithm reaches accuracy results within 5% of accuracy values obtained for supervised training. The number of exons predicted for *D. melanogaster*, for example, is 19,926 at the first iteration and 50,526 at convergence. Significant jump in Sn values is observed at step 4 (or iteration 3), when the length distributions and phase dependency are turned on.

The state durations described by a uniform distribution at the algorithm initialization step change to a bell-like shape at the algorithm convergence which is nearly indistinguishable from the length distributions derived from a set of annotated genes. This is demonstrated in Figure 4.9 for *D. melanogaster*.

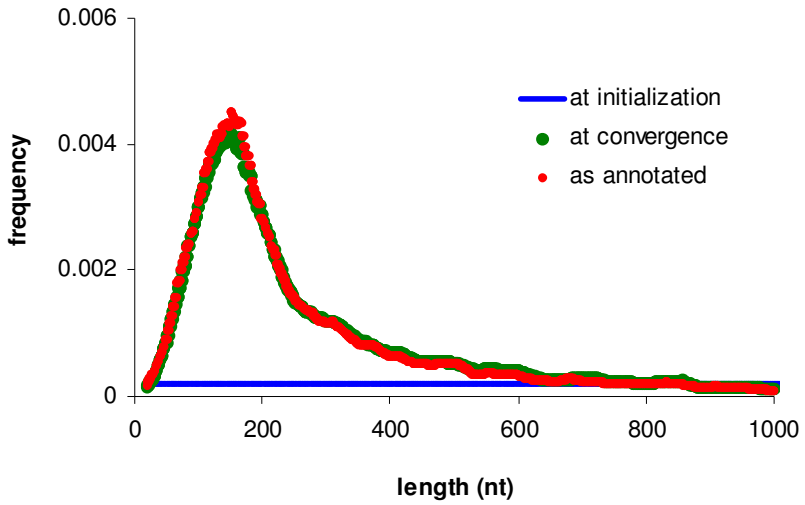


Figure 4.9 The shape of internal exon length distribution for *D. melanogaster* in iterations as determined at GeneMark-ES initialization and convergence as well as from the annotation. The length distribution of internal exons obtained from annotation and the algorithm convergence are nearly indistinguishable.

As the discrimination power of the models increases, the prediction accuracy grows in iterations as well. The splice site models which contain the minimum amount of information represented by corresponding dinucleotides at the initialization step are significantly enriched after the first iteration (Figure 4.10). Motif logos and the values of KL distance (in bits) between the motif and the background sequence show even further improvement achieved with at the algorithm convergence. In iterations Kullback-Leibler distance between the model parameters of protein coding and non-coding states grows (Figure 4.11) and decreases between models obtained from supervised and unsupervised trainings (Figure 4.12).

Interestingly, relative entropy values of nearly 0.4 bits between these states are sufficient for accurate gene prediction in prokaryotic species (given of course that model parameters for other elements such as RBS site model, state durations, stop signal models and such are well defined). For *A. thaliana* and *C. elegans* the drop in KL distance between coding and non-coding models (Figure 4.11) coincides with drop in accuracy in the last iterations (Figure 4.8). The decrease in Sp at the convergence is partially related to the fact that the genes in the test set contain flanking region that may carry coding sequences coming from incomplete genes. As the models become more sensitive, these partial regions are being better detected. Moreover, EM types of algorithms such as GeneMark-ES by their nature do not guarantee convergence at the global maximum of the likelihood function. Hence, the model at the convergence does not necessarily produce the best result but the one that is in the vicinity of the highest accuracy values.

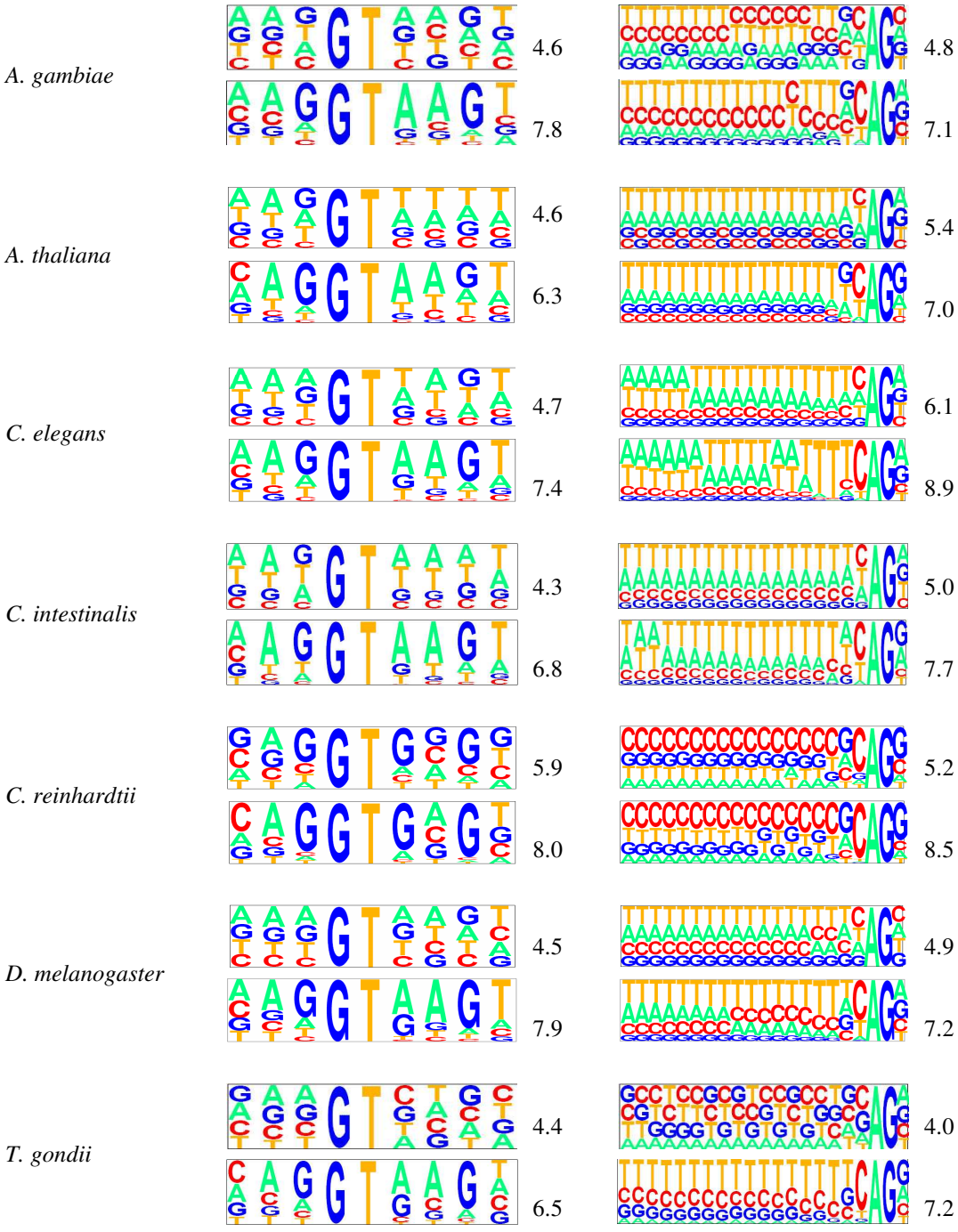


Figure 4.10 The splice site motifs (the donor site the left column and the acceptor site the right column) derived after the first iteration (top panel of panel pairs), and at the algorithm convergence (the bottom panel). First order KL distance in bits is shown next to the pictograms. The pictograms were obtained by using the software utility available at genes.mit.edu/pictogram.html.

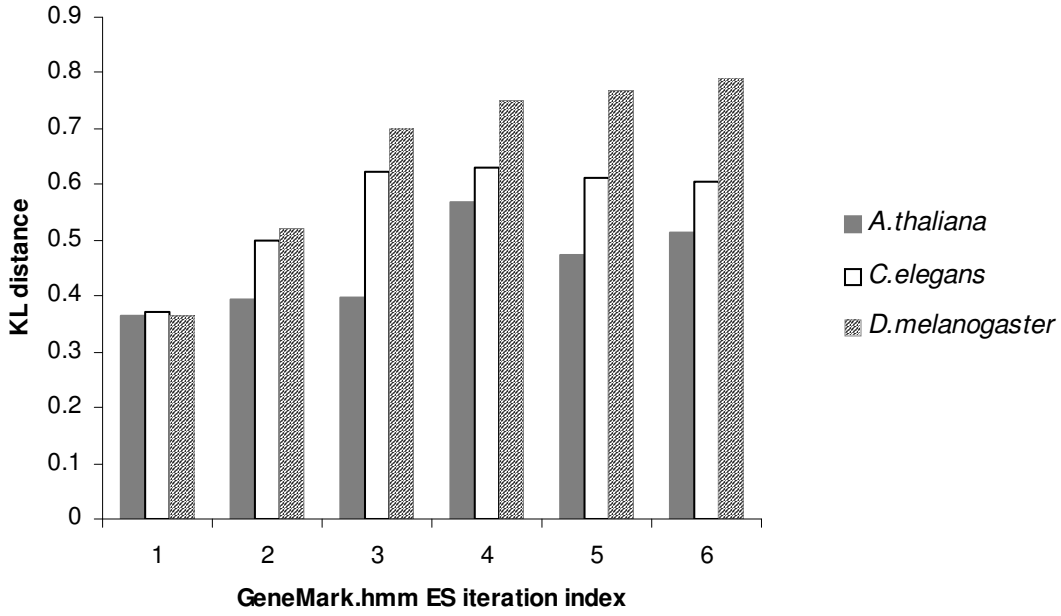


Figure 4.11 The KL distance (in bits) between models of protein coding and non coding regions shows significant growth in iterations. The decrease in KL distance at the last iterations observed for *A. thaliana* and *C. elegans* is in agreement with the results shown in Figure 4.8 where the Sp values for initial and terminal sites at iteration index 5 and 6 decrease as well.

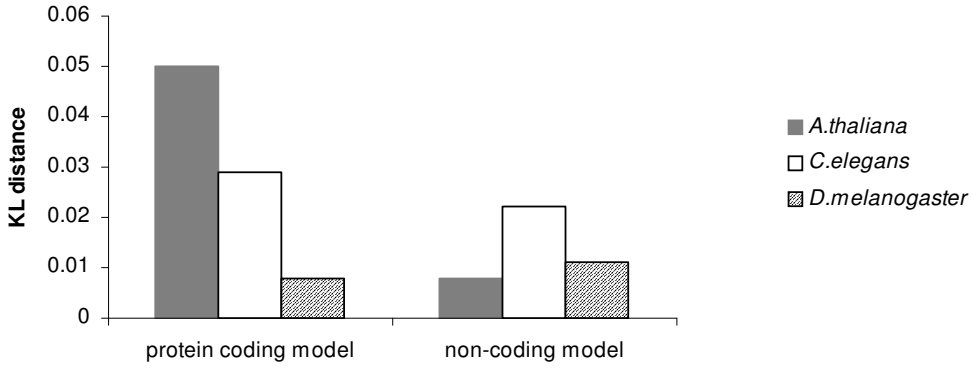


Figure 4.12 The KL distance (in bits) reflecting the divergence of models derived by supervised and unsupervised trainings. The KL values in this category are on the order of magnitude smaller than the KL distance observed between coding and non-coding (Figure 4.11).

The algorithm demonstrates interesting results when applied to C. intestinalis. Self-training identified two peaks in C. intestinalis localized at 60 and 300 nucleotides (Figure 4.13). These findings are consistent with results obtained by EST to DNA alignments (Dehal, Satou et al. 2002). Similar two-peak intron distribution is observed for other species representative of Ciona genus C. savignyi. Although shape of *T. gondii* intron distribution is similar to that of C. intestinalis the small peak is shifted toward even shorter lengths (28-45nt) placing splice site signal at a very close proximity.

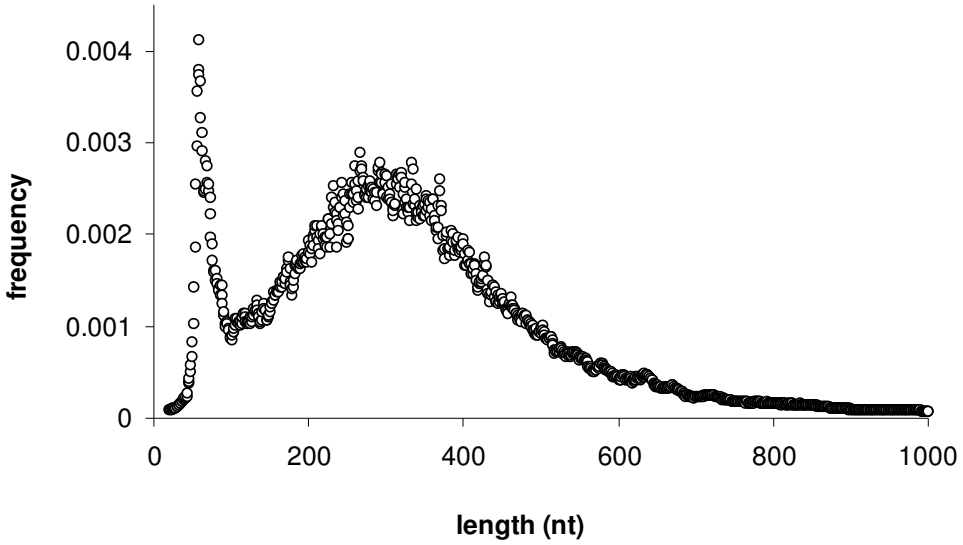


Figure 4.13 The intron length distribution for *C. intestinalis* obtained at the algorithm convergence shows non-unimodal distribution This shape is not typical for other species. Similar two-peak intron length distribution is exhibited by *Ciona savignyi*, a larger, 180 Mb relative.

4.3.3. Comparison with SNAP

The SNAP (Korf 2004), *ab initio* gene finding algorithm, is designed specifically for novel eukaryotic genomes with limited training set. SNAP develops a bootstrapped gene model by first choosing from already annotated and well-studied genomes. In the

selection process (made by user) it is recommended to use a model of those species that fall into close evolutionary distance from the genome in question. Next the model is employed to predict genes in genomic sequence. The set of predictions are then used to train the bootstrapped model. As an option, SNAP allows a choice of more than one model to parse the DNA sequence. In this case, the predictions of several models are combined to produce a training set for the bootstrap model.

Two test sets are used in this comparison. The first is the test set derived for this study and described in Chapter 3. The second is the set of sequences available at the SNAP program website (www.biomedcentral.com/content/supplementary/1471-2105-5-59-S1.gz).

While SNAP's supervised models for five eukaryotic genomes are available for download, the bootstrap model derivation procedure for SNAP is not included in the distribution package for download. In comparison with the bootstrap models only nucleotide accuracies are used as they are cited in the original publication (Korf 2004). For comparison purposes SNAP's bootstrap model that shows the best results for a given species is used to report the accuracies. For example, the bootstrap model based on *C. elegans* predictions is used to report prediction accuracy for *A. thaliana* since it performs better than other bootstrap models. Supervised models for SNAP are used to assess prediction accuracy in other categories of gene elements. Reportedly, SNAP performed better with a native model derived by supervised training than any of the bootstrap models (Korf 2004). Therefore, the comparison results with supervised models would provide a reasonable point of reference for algorithms performance.

Table 4.4 shows the accuracy results (Sn/Sp and average) for GeneMark-ES and SNAP's supervised models applied to the Test Set I described in Chapter 3. For the genomes used in comparison models derived by unsupervised training algorithm GeneMark-ES outperforms SNAP's supervised models in all categories in terms of (Sn+Sp)/2 values. The comparison results are different for well-studied (Table 4.4a) and novel species (Table 4.4b). For well-studied species the gap in average prediction accuracy on nucleotide level reaches 1.8% (*A. thaliana*). The positive difference in prediction accuracy increases in other categories. GeneMark-ES shows 5.5%, 5.6% and 5.8% higher results predicting internal exons in the test sets of *D. melanogaster*, *C. elegans* and *A. thaliana* respectively. The accuracy evaluation on exon level is important since mislabeling the exon boundaries will change the protein product of the predicted gene.

For the novel genomes the gap is greater across nearly all categories which is not surprising given the limited data available for the training set. The results are even more dramatic for *C. intestinalis* where GeneMark-ES outperforms SNAP by 12.0% and 22.2% in nucleotide and internal exon prediction accuracies. The effect observed with the novel species underlines the significance of the self-training algorithm which can be effective tool for gene prediction in novel genome.

The prediction accuracy results for GeneMark.hmm E, GENSCAN, GeneFinder and AUGUSTUS trained for *A. thaliana*, *C. elegans* and *D. melanogaster* have been reported for the test set obtained by Korf 2004 and are shown in Table 4.5a. GeneMark-E employing models derived by supervised training shows results comparable or better than the accuracy of other supervised gene predictors.

Table 4.4 Comparison of prediction accuracy results of GeneMark-ES and SNAP in terms of Sn and Sp values and their average. Bold font shows the larger value out of the two in corresponding category between GeneMark-ES and SNAP.

a) accuracy comparison for well-studied species

		A. thaliana				*C. elegans*				*D. melanogaster*			
		GeneMark-ES		SNAP supervised		GeneMark-ES		SNAP supervised		GeneMark-ES		SNAP supervised	
Internal exon	Sn	**91.2**	**89.5**	79.7	83.8	**94.0**	**92.7**	87.1	87.1	**91.3**	**90.5**	85.8	85.5
	Sp	**87.8**		87.8		**91.3**		87.1		**89.7**		85.2	
Donor	Sn	**94.0**	**92.2**	83.7	86.9	**96.2**	**93.5**	90.1	89.0	**92.8**	**90.0**	86.9	86.6
	Sp	**90.3**		90.0		**90.8**		87.8		**87.2**		86.2	
Acceptor	Sn	**94.0**	**92.1**	84.6	87.8	**97.3**	**94.5**	93.5	92.1	**93.0**	**90.0**	87.7	87.3
	Sp	**90.2**		91.0		**91.6**		90.6		**87.0**		86.8	
Initiation site	Sn	**80.1**	**78.3**	75.6	75.0	**85.8**	**77.4**	73.2	67.4	**83.9**	**78.7**	78.1	77.7
	Sp	**76.5**		74.3		**68.9**		61.5		**73.5**		77.3	
Termination site	Sn	**87.5**	**85.3**	84.0	83.5	**95.1**	**85.2**	89.4	80.9	**89.2**	**83.2**	78.1	77.5
	Sp	**83.1**		82.9		**75.3**		72.4		**77.2**		76.8	
Nucleotide	Sn	**97.7**	**96.3**	93.6	94.5	**99.1**	**96.4**	97.2	95.7	**97.9**	**95.4**	94.8	93.9
	Sp	**94.8**		95.3		**93.6**		94.1		**92.9**		92.9	

b) accuracy comparison for novel species

		A. gambiae				*C. intestinalis*			
		GeneMark-ES		SNAP supervised		GeneMark-ES		SNAP supervised	
Internal exon	Sn	**89.3**	**88.9**	81.7	84.7	**94.8**	**93.5**	80.9	71.3
	Sp	**88.4**		87.6		**92.1**		61.7	
Donor	Sn	**89.7**	**86.9**	80.8	83.1	**95.3**	**92.5**	82.7	72.6
	Sp	**84.1**		85.3		**89.7**		62.4	
Acceptor	Sn	**92.3**	**88.5**	83.3	83.7	**96.3**	**93.3**	83.6	73.5
	Sp	**84.7**		84.1		**90.3**		63.3	
Initiation site	Sn	**77.8**	**72.9**	65.2	66.2	**79.6**	**71.3**	61.0	52.2
	Sp	**67.9**		67.2		**63.0**		43.4	
Termination site	Sn	**86.1**	**78.9**	78.8	76.0	**85.4**	**75.9**	63.3	54.6
	Sp	**71.7**		73.2		**66.3**		45.9	
Nucleotide	Sn	**96.0**	**90.5**	87.6	84.5	**98.3**	**94.2**	90.1	82.2
	Sp	**85.0**		81.4		**90.0**		74.3	

Note: the results are based on test set Type I.

As it is stated above in comparison of the performance of the GeneMark-ES against the SNAP bootstrap algorithm the best out of seven possible bootstrap models for the given species is used. Table 4.5b shows that the average Sn/Sp values are higher for GeneMark-ES by 1.7% for *A. thaliana* (where the bootstrapped model is derived from the training set obtained from *C. elegans* model predictions), by 3.3% for *C. elegans* (where the bootstrapped model is based on combined predictions of *A. thaliana* and *O. sativa* models), and by 0.4% for *D. melanogaster* (where the bootstrapped model is based on prediction of *O. sativa* model).

Table 4.5 Comparison of prediction accuracy results of different gene prediction algorithms reflected in nucleotide Sn and Sp values and their average. Bold font shows the largest value for a given species.

a) Comparison of supervised algorithms.

		GeneMark-E		SNAP supervised		GenScan supervised		Genefinder supervised		Augustus supervised	
A. thaliana	Sn	**98.4**	**96.3**	97.1	96.2	79.9	86.4	-	-	-	-
	Sp	94.2		**95.2**		92.9		-		-	
C. elegans	Sn	97.7	**97.0**	97.6	95.9	-	-	**98.1**	96.7	-	-
	Sp	**96.2**		94.2		-		95.3		-	
D. melanogaster	Sn	93.2	**90.5**	**94.3**	90.4	-	-	-	-	92.4	**90.5**
	Sp	87.7		86.5		-		-		**88.6**	

b) Comparison of GeneMark-ES and SNAP bootstrap models.

		GeneMark-ES		SNAP bootstrap	
A. thaliana	Sn	**98.3**	**96.5**	96.6	94.9
	Sp	**94.7**		93.2	
C. elegans	Sn	**99.1**	**97.1**	96.7	93.9
	Sp	**95.1**		91.1	
D. melanogaster	Sn	**93.8**	**90.0**	92.5	89.6
	Sp	86.1		**86.6**	

Note: the results are based on the test set derived by Korf, 2004.

4.3.4. Minimum Genome Size for Successful Self-Training

Model parameterization for GeneMark.hmm employing unsupervised training shows consistently satisfactory results reflected in the algorithm accuracy evaluation on available test sets (Tables 4.1-4.5). The practical use of the algorithm is for genomes that are in early stages of genome sequencing project when the availability of a validated set of genes to derive model parameters for gene finder via supervised training is limited. The valid question then would be: "What is the minimum size of the input sequence needed to obtain reliable gene predictions?" To address this question the following experiment is carried out. For a given genome the sequences of various lengths are randomly extracted from genomic DNA and are used as inputs for unsupervised training. In these experiments the minimum length of a coding sequence in filtering procedure is set to a less stringent value of 300 nt to provide sufficient size of training set and to avoid possible parameter overfitting in the course of iterations

Figure 4.14 shows average prediction accuracy in the category of internal exon prediction, characterized by the value of (Sn + Sp)/2, as a function of the input sequence length. For *A. thaliana* (87%), *C. elegans* (91%) and *D. melanogaster* (90%) the prediction accuracy is reasonably high for the input sequence size of 10 Mb. The number of iterations, up to the point of algorithm convergence, however, increase on average two-fold. Below 10 Mb the algorithm shows a sharp drop in prediction accuracy for all species. These results suggest that 10 Mb of sequence is sufficient for unsupervised parameterization of the statistical model (HSMM) for the GeneMark-ES gene finding algorithm. Several factors may influence this estimate. In low gene density genomes, for example (larger and more complex genomes such as vertibrates), and/or in species

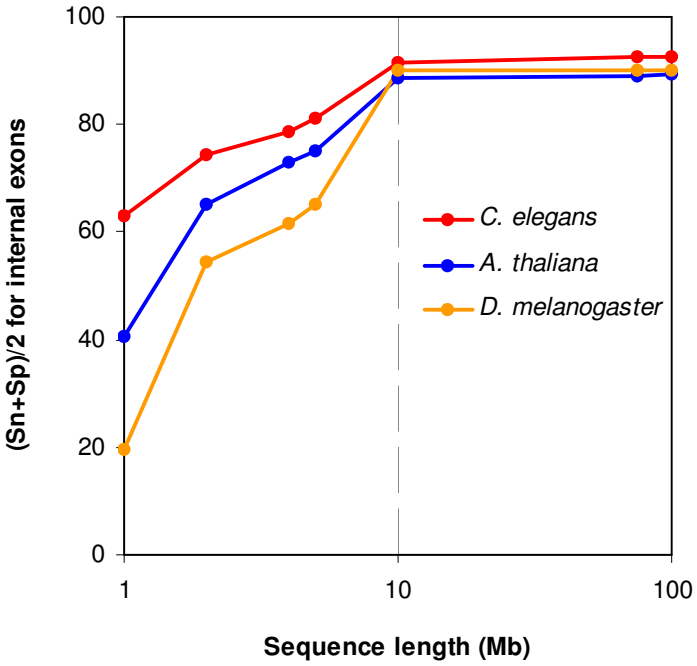


Figure 4.14 Dependence of the average prediction accuracy of internal exons, defined as average Sn and Sp on the input sequence length. The results suggest that GeneMark-ES with 10 Mb input data provides accurate models for gene prediction.

heavily populated with transposable elements, the size of the input sequence is expected to be higher. On the other side of the spectrum are genomes of low eukaryotes such as *S. cerevisiae* which may not be suitable species for this type of unsupervised training since they usually contain a relatively small number of introns (~200 for S. *cerevisiae*). The main difficulty in this case is the parameter estimation for state durations and splice sites (this problem is addressed in Chapter 6). Otherwise, for species with genome organization similar to those described in this chapter the unsupervised training with an

input sequence where the size is least 10 Mb will provide accurate models for gene prediction.

Usually for novel species in an early stage of genome sequencing, when the extrinsic evidence for a sufficiently large training set is unlikely to be found, GeneMark-ES, is perhaps the only automated gene prediction algorithm suitable to analyze the sequence data in hand.

As the sequencing project moves forward, a training set for supervised gene prediction becomes available and so does the possibility for an update of unsupervised training. Moreover, as shown above for well-studied genomes the models obtained by unsupervised training demonstrate superior or comparable results in comparison with models derived by supervised training.

4.4 Initialization Impact on Unsupervised Training

Three methods of parameter initialization for GeneMark-ES are implemented as discussed in Section 4.2.1. Regardless of the initial model the algorithm followed the same procedure with unchanged fixed settings and thresholds.

The comparison is done using the test sets of *A. thaliana*, *C. elegans* and *D. melanogaster*. The accuracy results indicate that the unsupervised training algorithm is stable and even with a weak initialization model it demonstrates satisfactory results. Figure 4.15 shows that while the results based on the ORF based initialization model performance is within close proximity (0.2%) of results with heuristic initialization. The G+C fixed value initialization shows higher variability and in the worst case

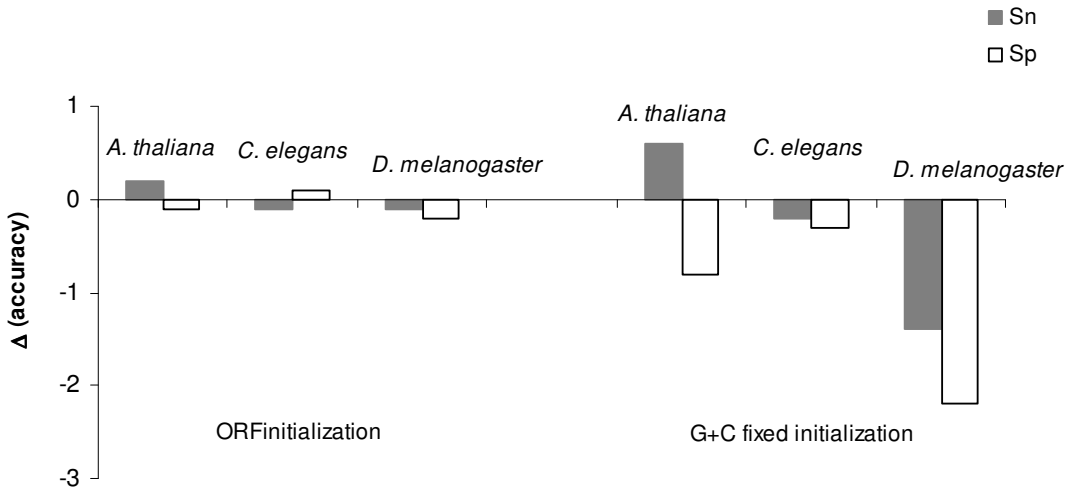


Figure 4.15 The difference in internal exon prediction accuracy values between GeneMark-ES initialized by heuristic models and two alternative approaches. Left: ORF initialization, right: G+C fixed initialization (see Section 4.2.1 for details). The positive values indicate better performance of the heuristic initialization. The results show that the algorithm is stable and is converging to biologically relevant point regardless the initialization models.

(*D. melanogaster)* results that are 1.4% (Sn) and 2.2% (Sp) lower than that of heuristic models.

In terms of convergence reflected by the number of iterations additional steps are needed for both initialization methods. When GeneMark-ES is initialized with these models it needs an additional iteration for *C. elegans* and two extra iterations for *D. melanogaster* when compared to initialization with a heuristic model. In the case of *A. thaliana* it needs to run six and seven iterations when initialized by ORFs and fixed G+C models respectively before it reaches convergence (vs. the five iterations with heuristic starting point).

Based on these results a question may be posed: "Given a well performing supervised model can it (the supervised model) effectively be used as an initialization point?" To address this question the supervised models for *A. thaliana*, *C. elegans* and *D. melanogaster* are plugged into the algorithm as initial models. Since the supervised models are relatively well-trained the parameter re-estimation step is done in full mode (see Sections 4.4 and 4.5 for details). In general the results are not significantly different from that of the unsupervised training. When applied to *C. elegans*, for example, the algorithm starts with the model performing with 90.9% (Table 4.1) average accuracy for internal exons. After four iterations it converges to 92.7% which exactly coincides with that of unsupervised training with heuristic initialization (Table 4.1). Similar results are observed for *D. melanogaster* where the accuracy reaches 91.3%, close to the value obtained by unsupervised training (91.0%). In the case of *A. thaliana* the accuracy performance gap between supervised and unsupervised training is 0.4% (Table 4.1). The application of iterative training based on the supervised model in this case yields an average of 89.8% internal exon accuracy. These results suggest that with a given model architecture and the training process the algorithm convergence point is within close proximity regardless of the parameter initialization procedure.

4.5 Repetitive Elements in The Course of Unsupervised Training

Mobile elements containing long open reading frames (e.g. LINEs, ERV and MaLR) and simple repeats potentially can create problems for the training process especially when dealing with higher eukaryotes (vertebrates, mammals). Ideally, the

genomic sequences should be scanned and masked from repetitive elements before applying the self-training algorithm. In reality, however, this task is far from trivial.

Repetitive elements of *A. thaliana*, *C. elegans and D. melanogaster* occupy significantly smaller portion of their host genome compared to that of higher eukaryotes. In this group the simple repeats are the most frequent and diverse group in *D. melanogaster* (Katti, Ranjekar et al. 2001). *A. thaliana* genome however contains the highest number (5,500) of transposable elements representing 10% of its genome (Initiative: 2000). For most of the species presented in this thesis already masked genomic sequences are publicly available for download. GeneMark-ES is run on these sequences with *a priori* knowledge of repeat coordinates. The algorithm is restricted to predicting protein coding genes in masked sequences. Also, repetitive regions are not considered in the parameter estimation step. The difference in prediction accuracy of internal exons for the algorithm employed with and without masking of repetitive sequences is shown in Figure 4.16.

As stated above the masking should be taken cautiously as the algorithm performance shows mixed results. Repeat masking for *C. elegans* results in positive growth of both Sn (0.8) and Sp (1.4). For *C. intestinalis*, marginal improvement is observed for Sn (0.1). For other genomes Sn and Sp show a decrease with masked input sequence. The Sp values are more affected by the masking procedure than the Sn regardless of the positive or negative outcome. At first surprising, this trend can nevertheless be explained. DNA sequence masking usually involves use of several software packages. RepeatMasker (Smit, R., Green, H., Green P. unpublished work)

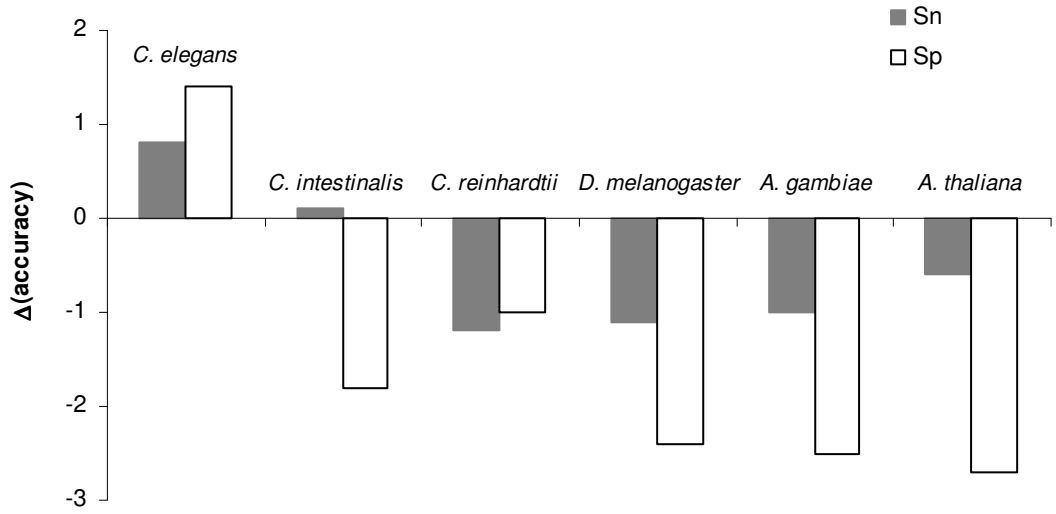


Figure 4.16 The difference in internal exon prediction accuracy values between GeneMark-ES run on unmasked and masked sequences. Positive values indicate better performance of the GeneMark-ES with unmasked sequences.

identifies the repeats by scanning the nucleotide sequences against the library of repetitive elements.

TRF developed by Benson (Benson 1999) is an effective approach for tandem repeat finding. Each of these approaches completes the task with its own error rate. One possible scenario for the ineffective masking is partial identification of a particular repetitive element in DNA, e.g. only part of ORF within the mobile element is masked; this is possible due to the differences between the genomic sequence and the sequence in the repeat library.

The training process is also negatively affected when the masked sequence *A* contains partial regions from both coding and non-coding states. In both of these situations gene finder will predict the exposed (unmasked) segment of the ORF or coding region and given the restrictions imposed by masking. Frequently it will be forced to

complete the parse by making a false prediction elsewhere in the sequence. Therefore, when working with a masked sequence an extra step possessing predictions should be developed. Depending on species and masking procedure similar effects are observed for supervised parameterization (Shmeleva N. and Lomsadze, personal communication 2004).

4.6 Novel Genes Identified by GeneMark-ES

In the preceding sections the algorithm performance was shown on an already known set of genes. Another valuable characteristic of the algorithm is its ability to find new genes. Newly identified and biologically interesting genes are discussed in this section. The procedure of novel gene search described in Chapter 3 employs similarity search between sets of predicted and annotated gene products. The subset of predicted proteins with no hit to annotated proteins is used to search against NR and CDD. The proteins that satisfying the filtering thresholds are selected into the set of novel genes. The most interesting findings are shown in Table A1 in Appendix. The full list of novel genes is available at http://nar.oxfordjournals.org/cgi/data/33/20/6494/DC1/1. The sequences of these genes in GenBank format are available at http://nar.oxfordjournals.org/ cgi/data/33/20/6494/DC1/2.

For *C. elegans* the wormbase provides gene predictions of TWINSCAN and GeneFinder which allow the direct comparison with GeneMark-ES. Generally the exon-intron structure of the predicted genes are in agreement among these programs. An example of missed *C. elegans* gene by two other gene prediction programs is shown in

Figure 4.17. The gene product has significant whole length similarity (*e-value* 2e-29) to a hypothetical protein in *C. briggsae*.

4.6.1. New "Housekeeping" or Important Metabolic Genes

The following predicted gene products are anticipated to be found in the proteome of genomes under study but are missing from the annotated sets of proteins.

In *A. gambiae* (Table A1b):

- Gene containing cytochrome c oxidase subunit VIc (#2) a component of the mitochondrial electron transport chain which involved in catalysis of O_2 reduction and pumps protons across the membrane.
- Nup84p (#21); nuclear pore complexes (NPCs) are essential components for RNA export (Rollenhagen, Hodge et al. 2004). Transport proteins required for mRNA export (*S. cerevisiae*) are found in several complexes including Nup84p.

In *C. intestinalis* (Table A1d):

- Mitochondrial ribosomal protein L10 homolog (#6). In eukaryotes a protein complex consisting of L10 and number of multiple copies of L12(e) interacts with 39S ribosomal subunit.
- Single exon gene product with a hit to RNA polymerase I associated factor 53 from human, mouse and Xenopus (#23). This protein also shows a similarity to the hypothetical protein LOC395052 from Xenopus, suggesting that a Xenopus protein may also play a role in translation;

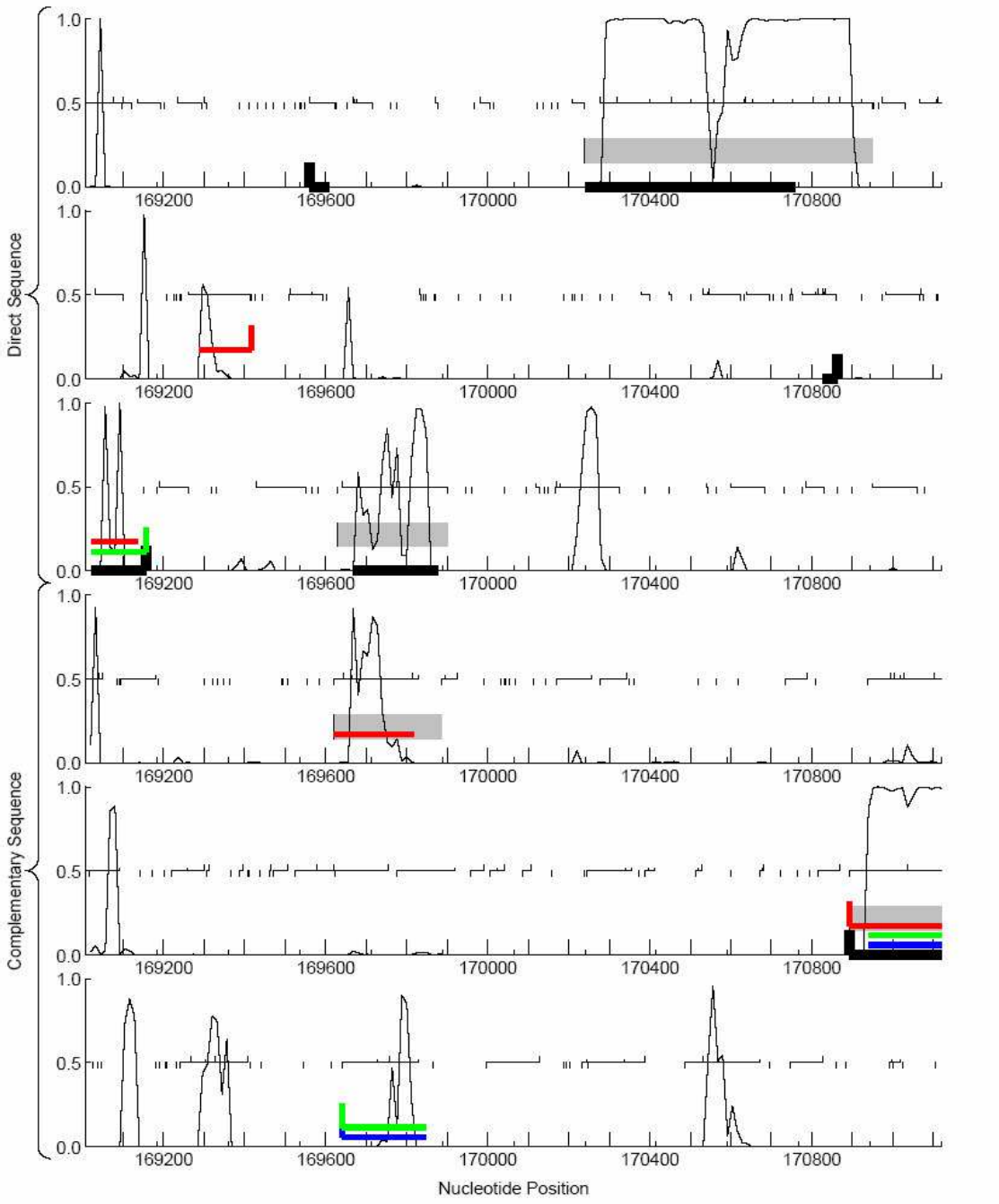


Figure 4.17 Newly predicted *C. elegans* gene by GeneMark-ES (black) which is missed by Twinscan (green) Genefinder (red). As GeneMark graph shows the two internal exons exhibit high coding potential. The predicted gene product shows significant similarity closely related species *C. briggsae*.

- Two-exon gene producing protein of length 764 aa which has a similarity to translation initiation factor 3 and subunit 8 from rat (#27).

In *C. reinhardtii* (Table A1e):

- Homologs of the ribosomal proteins S21e (#1), S21 (#6), S9/S16 (#10);
- Nucleolar protein Nop10p (#2) essential factor in eukaryotic 18S rRNA ribosome biogenesis.
- A four-exon gene (#4) containing a region highly similar to protein that is a component of the Sec61 protein secretory system, studied previously in yeast and also found in humans and apes; Sec61 was recently suggested to be involved in endoplasmic reticulum (ER) associated degradation pathway (Scott and Schekman 2008).

In *D. melanogaster* (Table A1f)

TTD-A (#1) gene recently found to be directly responsible for TFIIH transcription complex stability in humans (Giglia-Mari, Miquel et al. 2006); it is also an ortholog of transcription factor TBF5 of *Gallus gallus*). TTD-A is involved in general control of transcription and transcription-associated DNA repair, and possibly in cell cycle regulation. Mutations in this protein cause trichothiodystrophy. TTD-A presence in *Drosophila*, a model organism which is well-studied genetically, opens a new opportunity for further functional analysis of TTD-A.

4.6.2. Genes with Homologs In Phylogenetically Closely Related Organisms

While the presence of these genes is expected based on the phylogenetic positions of the given organisms, they have not been identified previously by other methods.

In *A. gambiae* (Table A1b)

- Gene that encodes a homolog of 'royal jelly' protein in an *A. gambiae* (#23) involved in control of cast differentiation in honey bee. It is also homologous to Drosophila protein CG7463-PA. Presence of such a protein in *Anopheles* provides additional support as well as insights to its phylogeny among different groups of insects.
- Homolog of the mammalian male enhanced antigen 1, suggested to play an important role in the late stage of mammalian spermatogenesis (#23), is another example from *A. gambiae*. This predicted gene product is also found in *Drosophila* (CG14341-PB). Identification of this protein in mosquito and its highly similar hit in Drosophila further confirms its broad evolutionary conservation, despite the possible divergence of its specific biological roles.

4.6.3. Unexpected Genes

In this section genes whose presence in these species were not reported in annotation and generally are not expected to be found in the host genome are listed. Discovery of these genes provides new insights to the evolution of the specific gene families and/or biology of the specific organisms. The following list of genes belongs to this group.

In *A. gambiae* (Table A1b):

Gene #19 a 2,202 nt long ORF, a homolog of the mammalian neurochondrin. This protein is produced in bone-producing cells, mammalian chrondrocytes and several neurons. Neurochondrin is thought to play a specific role in bone metabolism, neuron growth, chondrocyte differentiation and cell resorption regulation (Ishiduka, Mochizuki et al. 1999, Dateki, Horii et al. 2005). Presence of this protein in species which do not possess an internal skeleton suggests that the neurochondrin may have broader functions than initially thought. Homologous to this protein is also found in *Drosophila*'s CG2330-PA. To our knowledge, similarity between these proteins and neurochondrin has not been recorded previously.

In *C. elegans* (Table A1c):

Tetracycline resistance protein of group C (#6) is known to be present in prokaryotes such as Shigella and in the transposon *Tn10*. Its existence in eukaryotic species may suggest a lateral (horizontal) gene transfer. Horizontal transmission between bacteria and nematodes provides a new insight into biology of these organisms.

4.7 Why Do the DNA-To-Protein Searches Miss These Genes?

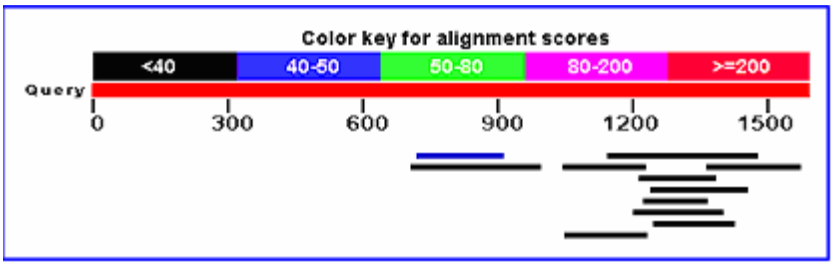
Gene identification by DNA-to-protein search is regularly performed by annotation groups. As discussed in Chapter 2 gene identification methods that employ similarity searches are limited in their ability to reconstruct gene structure from the alignment; the level of difficulty increases proportionally to the evolutionary distance between the hosts of the sequence in question and the protein in the database.

In the genomic sequence of *C. intestinalis* for example GeneMark-ES finds five-exon gene producing a protein with length of 244 aa (Table A1d). Similarity search against NR database and CDD returns significant hit (e-value of 10^{-16}, over 70% identity over the whole length) to *L10* protein in NR which contains conserved ribosomal domain RpIJ. Applying *blastx* (Gish and States 1993) to the transcript, however does not return practical sensible results (Figure 4.18a). Translated *blast* search shows better results for DNA sequence of *C. reinhardtii*'s novel 5-exon gene (Table A1e #10). While presence of multi exon gene (Figure 4.18b) is observed out of five exons consensus gene structure of alignments supports only two internal exons.

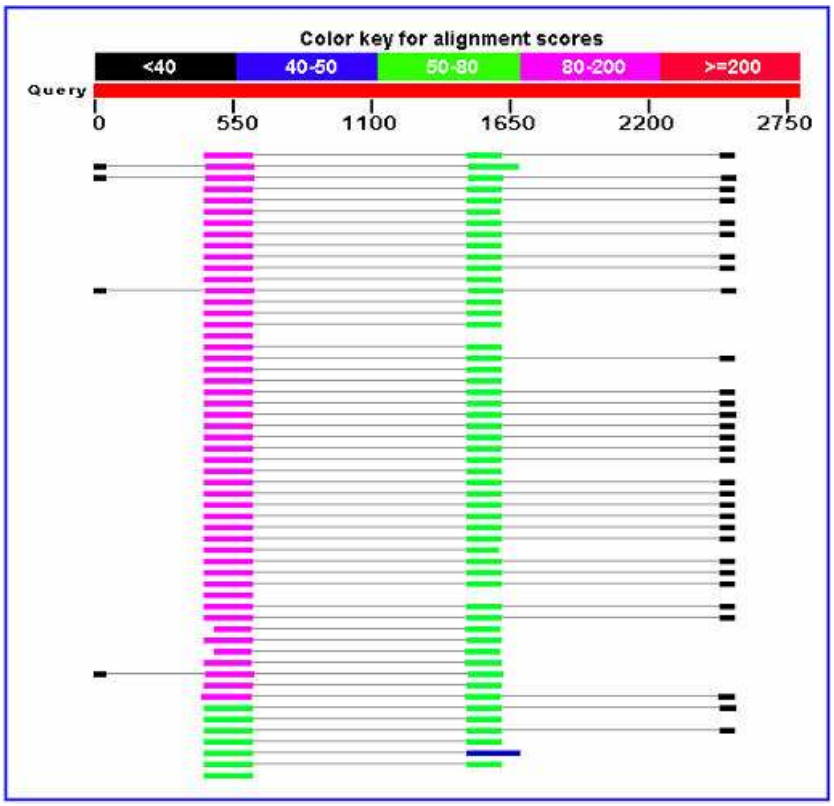
4.8 Conclusions

Eukaryotic gene finding follows similar development that prokaryotic gene prediction experienced in the past. Initially, prokaryotic predictors entirely relied on the training set eventually evolved into algorithms utilizing unsupervised training methods sequencing projects demanded urgent development of an automatic *ab initio* gene finding algorithm (Delcher, Harmon et al. 1999, Baldi 2000, Besemer, Lomsadze et al. 2001). The acceleration of genome sequencing process and consequently the significant increase in the number of eukaryotic

GeneMark-ES, a self-training algorithm for eukaryotic gene prediction is described in this chapter. The proposed method represents a novel approach in HMM model parameterization from anonymous DNA sequence. Algorithm is particularity useful for eukaryotic genomes being in the initial stages of sequencing. Accuracy evaluation on the sets of validated genes shows that GeneMark-ES performs equally well or better than



a) *L10* in *C. intestinalis*



(b) *R16* homolog in *C. reinhardtii*.

Figure 4.18 Snapshot of *blastx* results for unspliced DNA sequence of newly identified gene products. The *blastx* search does not return reliable similarity hits to reconstruct the gene structure.

other gene prediction algorithms. Self-training algorithm with input sequence of 10 Mb is shown produce reliable gene models reasonable accuracy results. The algorithm demonstrates stable converging properties with respect to different initialization models.

Gene prediction in fungi and in low eukaryotes is addressed in Chapters 6 and 7, respectively. Several complications potentially are possible in both low and higher eukaryotes which are related to their genome organization. In low eukaryotes, as the genome size and number of introns per gene decreases, the derivation of accurate model parameters becomes difficult. Furthermore, the contribution of donor and acceptor sites into splicing mechanism is relatively small which is balanced by well conserved branch point motif. For the higher eukaryotes whose genome composition is inhomogeneous and densely populated with transposable elements additional steps are necessary to be taken into consideration (see Chapter 7).

CHAPTER 5

SELF-TRAINING ALGORITHM GeneMark-ES-2

FOR FUNGAL GENE FINDING

5.1 Introduction

Eukaryotes represent a wide array of species that vary in their genome organization and complexity creating unprecedented challenges for automatic gene prediction programs. Currently more than 300 fungal genome projects from seven phylum are reported (http://www.genomesonline.org and (Liolios, Tavernarakis et al. 2006). Most of these eukaryotic species contain a significant amount of information required for intron splicing in branch point (BP) motif. Lim et al. (Lim and Burge 2001) have shown that while BP contribution to intron splicing is less than 5% in *D. melanogaster, A. thalilana, C. elegans,* and *Homo sapiens* its input for *S. cerevisiae* is about 40%. For well-studied genomes of *D. melanogaster*, *C. elegans*, and *A. thaliana*, the performance of the self-training algorithm, GeneMark-ES was shown to perform with matching or better accuracy of gene prediction than algorithms which employ supervised training in parameter estimation procedure (see Chapter 4 and Lomsadze, Ter-Hovhannisyan et al. 2005). The GeneMark-ES architecture does include the models for BP motif since for these species its contribution for accurate splicing is marginal (Lim and Burge 2001). A more complex intron model however is necessary for genomes where introns contain conserved branch point. This chapter describes the *ab initio* gene finding algorithm GeneMark-ES-2, an extension of GeneMark-ES, specifically designed

for fungal genomes. It employs a new intron submodel to better reflect the fungal gene organization, and it accounts for genes without consensus BP motif.

Gene prediction from an anonymous DNA sequence is carried out in parallel with statistical model parameterization and applied to species from *phylum* of Ascomycota, Basidiomycota and Zygomycota shown in Figure 5.1. The figure largely reflects the distribution of phylum that is seen at http://www.genomesonline.org, where Ascomycota species have received the greatest attention with 243 genomic projects in progress; the trend is not surprising given the impact this category has on agriculture and the food industry. The common ancestor for this group according to different estimates, is placed from 600-1,800 million years ago (Redecker, Kodner et al. 2000, Taylor and Berbee 2006) providing sufficient time for genome evolution. Hence, the observed variation in exon-intron structure is not surprising (Figure 5.2); while *S. pombe* shows an average of one intron per gene *C. neoformans* exhibits number nearly five times greater. While the average intron length does not show significant a dependence on the average number of introns per gene, the average exon length is negatively correlated with this number (Figure 5.2). This dependency is not surprising given the limited "room" provided by the gene for the introns to be inserted. For example, if a gene contains two exons each with length L and an intron is inserted into one of these exons then on average the exon length reduces to 2L/3. The same relationship is true in the case of intron deletion. In this case, however, the average length increases from 2L/3 to L. This trend is also observed for *A. thaliana, Oryza sativa* and *Homo sapiens* genes (Atambayeva 2008).

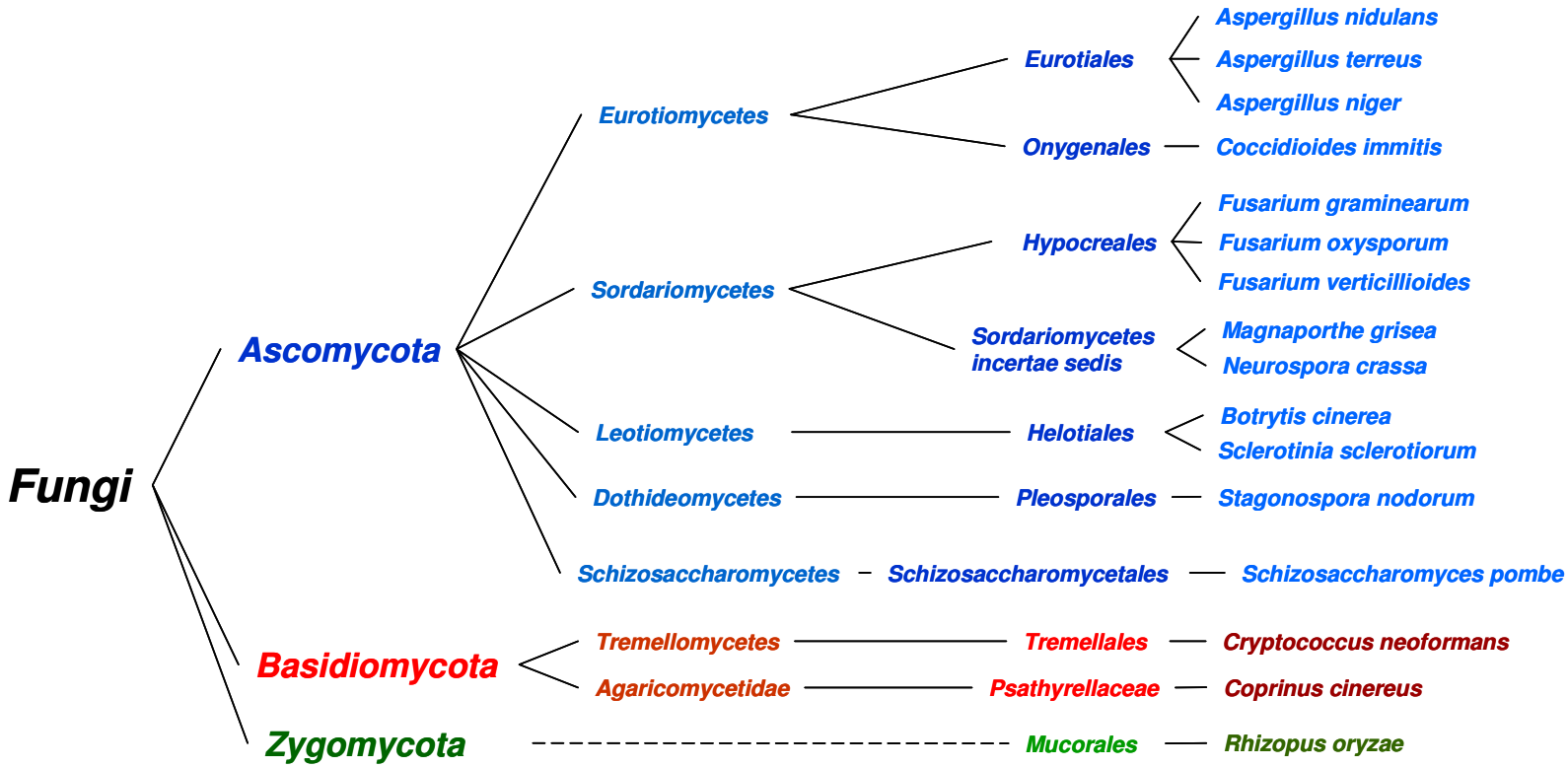


Figure 5.1 Phylogenetic relationships of the fungal species under consideration (source: http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy).

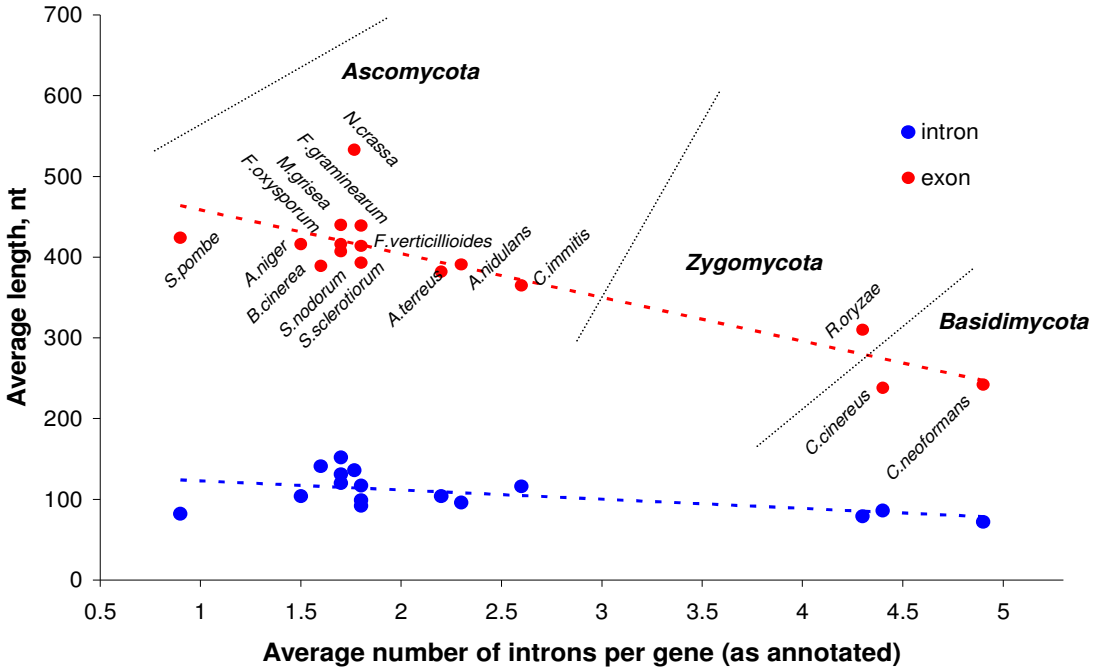


Figure 5.2 Variability of gene organization in fungal genomes.

In contrast to higher eukaryotes, fungal genomes, in addition to having well defined BP, are smaller in size as well as in the relative amount of the non-coding sequences ranging from approximately 50% in *B. cineria* to 70% in *N. crassa*.

The introduction of a new intron submodel in GeneMark-ES-2 and its application to these species leads to a significant increase in gene prediction accuracy compared to other approaches. The results also indicate that several of these genomes are possibly over-annotated and some are under-annotated. The algorithm is able to detect biologically important genes which currently are missing in the annotation.

Presently GeneMark-ES-2 is employed as part of annotation process at the Broad Institute, University of Hawaii and the Joint Genome Institute.

5.2 Methods

The iterative approach in deriving model parameters for GeneMark-ES-2 starts by following the path of GeneMark-ES (Sections 4.2.1 - 4.2.4). To introduce more complex intron model GeneMark-ES-2 carries significant changes in parameter estimation process and HMM architecture. With integrated modifications GeneMark-ES-2 continues the iteration until it reaches the point of algorithm convergence.

5.2.1. Changes in HMM Architecture

An enhanced intron model provides two alternative paths of hidden states for an intron sequence (Figure 5.3). The lower path consists of "intron" and "long acceptor" states as it is also employed in GeneMark-ES. The main purpose of the lower path is to make provisions for introns containing a weak (or challenging to detect) branch point signal. Four additional hidden states are presented in the upper path:

1. Upstream spacer (5' spacer)

This state generates a nucleotide sequence situated between donor and BP sites; nucleotide composition of this sequence and the sequence generated by the "intron" state in the bottom path are the same.

2. Branch point site

This state emits 9nt long nucleotide sequence generated by a positional Markov model of zero or first order. The order choice is based on the size of available data in the training set.

3. Downstream spacer (3' spacer)

The sequence between BP and acceptor site is modeled by the first order homogeneous Markov model and characterized by state durations. The nucleotide

composition of downstream spacer shows strong asymmetry as the frequency of thymine is over-represented in comparison with that of adenine.

4. Short acceptor

This state emits only two nucleotides upstream of canonical acceptor site "AG" which is suitable for fungal genomes whose introns do not possess a poly-pyrimidine (poly- Y) tail upstream of the acceptor site. Generally, introns with no poly-Y tail exhibit conserved BP site which is relatively easily to find by motif finders.

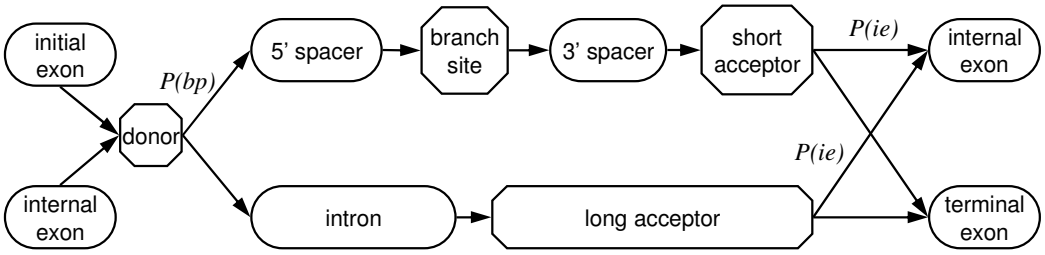


Figure 5.3 Hidden state diagram for the enhanced intron model (the diagram is shown for the direct strand only).

5.2.2. Changes in the Process of Unsupervised Parameterization

As stated above the iterative unsupervised training procedure follows the path of the original algorithm (see Section 4.2) up to the end of the 4^{th} iteration. Using the training set obtained from the predictions that are based on the model derived in the previous step, the algorithm sets "free" the parameters associated with state durations and phases of introns. At this point the new BP mode is activated by determining the model parameters for the states of the new intron submodel as follows.

Aligned BP site motifs are necessary to derive the BP site positional frequency model. The difficulty however is that neither the training set nor the parse of the current

iteration provide such an alignment. To overcome this challenge Gibbs sampler algorithm (Lawrence, Altschul et al. 1993, Thompson, Rouchka et al. 2003) is utilized to align segments of intronic sequences as it is described in Section 5.2.3. Positions of BP sites identified by Gibbs sampler are used to determine the downstream spacer parameters i.e. first order Markov chains and length distribution. The BP model is then used to scan the upstream region of introns which were not considered for the alignment step and to identify the highest scoring motifs in this remaining set. The positions of the BP in the whole set of putative BP sites are then used to determine state duration of the upstream spacer for which the compositional model of introns (usually of the 5^{th} Markov order) is used to describe the emission probabilities.

Remaining parameters that are needed to be addressed are the state transition probabilities from donor site to the upper path *P(bp)* and from the short or long acceptor to the internal exon *P(ie)* (Figure 5.3). Similarly to the parameter initialization step these transition probabilities are assumed to be uniformly distributed. The state transition probabilities *P(bp)* and *P(ie)* are estimated in subsequent iterations. The value of *P(bp)* reflects the number of genes with evolutionary conserved BP sites. Therefore, after completion of the 5^{th} iteration with the model in which *P(bp)* is set to 0.5 the value of *P(bp)* is determined from the ratio of the number of introns predicted (emitted) by the hidden states of the upper path of the algorithm to the total number of introns.

The state transition probability from intron to internal exon *P(ie)* is related to the average number of exons per gene which exhibits properties of geometric distribution. Figure 5.4 displays counts of genes with different number of exons per gene The transition probability value of *P(ie)* is estimated as a parameter of geometric distribution

that is fitted to the distribution of the number of exons per gene observed in the training set at the current step iterations.

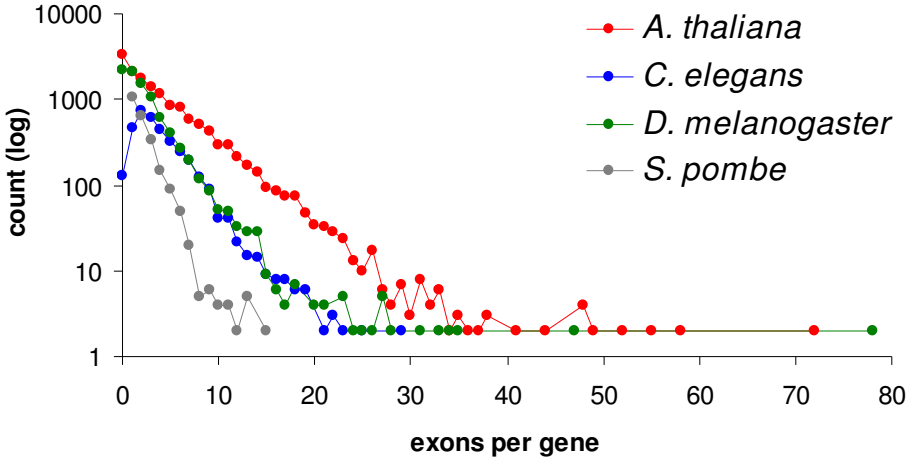


Figure 5.4 Counts (in log scale) of genes with different number of exons per gene as calculated from sets of genes confirmed by cDNA (*A .thaliana*, *C. elegans*), all annotated genes (*D. melanogaster*), and a set of genes with protein product showing full length similarity to a protein in the SwissProt database (*S. pombe*).

5.2.3. Gibbs Sampler. Overview and Settings

The Gibbs sampler algorithm is run in the *site sampler mode* which finds one motif per sequence. A detailed description of this approach and user manual can be found at http://bayesweb.wadsworth.org/web_help_text.Gibbs_versions.html.

The program is run without enforcing a constrains, e.g. fixing adenine in the BP site (Hebsgaard, Korning et al. 1996), masking sequence segments that do not contain BP to reduce the search space (Lim and Burge 2001), using information available from the related species to create initial starting point (Neverov 2003). Instead a subset of introns is selected form the training set to be used as an input to Gibbs sampler. Introns which

have length falling into the range of ± 10 nt from the maximum intron frequency are chosen into this test set. In addition, to decrease motif finder's running time the intron sequence is reduced to a segment of 50nt upstream of the acceptor site. There are two reasons for this pre-selection step. First, the error rate in short intron identification is higher. In fact, false predictions are likely to occur in the cases of both long and short introns due to the noise in intron length distribution caused by the relatively small number of points in regions outside of the maximum frequency of the intron. Second, the branch point in long sequences is located farther than 50nt upstream of acceptor site. In this case the segment with no BP motif is rather damaging the sampling procedure.

The algorithm is initialized by randomly assigning BP motif positions to each sequence and determining the model parameters of BP site and the background sequence. Then, it proceeds with consecutive steps of sampling and predictive update; the default number of 500 iterations is not changed in the runs. In order to avoid Gibbs sampler choice among more than one relatively strong motifs as the sampling algorithm reaches the near optimal results the positional frequency model is derived from those predicted motifs which are selected as motif start position within iterations at least 50% of time.

The default 10% pseudo count that is designed primarily for protein sequences is reduced to 0.1%. Moreover, this is required for more accurate estimates of nucleotide composition at the BP itself since the nucleotide "A" in the branching position is not fixed.

5.3 Results and Discussion

GeneMark-ES is applied to genomic sequences of fungal species from three different phyla Ascomycota, Basidiomycota and Zygomycota (Figure 5.1). The sixteen fungal species under consideration span over large evolutionally distances exhibiting significant variability in their genomic characteristics as shown in Table 5.1 and Figure 5.2. To better reflect the gene structure of the fungal genomes the algorithm employs the new intron model that provides conditions for genes that contain rather strong BP signal.

5.3.1. Algorithm Accuracy Evaluation

Species specific the test sets of Type I and Type II are generated as described in Section 3.4.1 to estimate GeneMark-ES-2 prediction accuracy in terms of Sn and Sp and their average. The set of artificial chromosomes are developed for *S. pombe* to assess the rates of gene splitting and merging as well as the frequency of gene prediction in the intergenic region.

Tables 5.2 and 5.3 show the algorithm accuracy results and its comparison to GeneMark-ES, the original unsupervised training algorithm as applied to test set Type I and Type II, respectively. For most of the species the algorithm shows significant improvement in accuracy results reflected in the difference in Sn and Sp (Figure 5.5; also Tables 5.2 and 5.3 columns δ). The best results in comparison to the original algorithm are obtained for *M. grisea* (13.3% in Sn and 6.1% in Sp) while only marginal increase is observed for *R. oryzae* (0.1% in Sn and 0.4% in Sp).

Table 5.1 Characteristics of the sixteen fungal genomes and the complements of predicted and annotated genes. Gene predictions were generated by the algorithm with enhanced intron submodel at the convergence point of self-training**.** Annotation data from EMBL (*A. niger*), NCBI (*S. pombe*) and Broad Institute (http://www.broad.mit. edu/) as of May 2008.

Species	estimated genome size (Mb)	GC content (%)	number of genes		number of single exon genes		number of introns per gene		number of introns per spliced gene	
			annotated	predicted	annotated	predicted	annotated	predicted	annotated	predicted
A. nidulans	31	50	10,701	10,445	1,453	2,278	2.3	2.0	2.7	2.6
A. niger	34	50	14,101	11,342	1,538	2,405	1.5	2.1	1.7	2.7
A. terreus	29	52	10,406	10,859	1,538	2,288	2.2	2.1	2.6	2.7
B. cinerea	26	43	16,448	11,890	4,316	2,624	1.6	1.8	2.2	2.3
C. immitis	29	46	10,457	8,435	1,449	1,903	2.5	2.0	2.9	2.6
C. cinereus	38	51	13,544	12,952	1,011	1,480	4.4	4.5	4.8	5.1
C. neoformans	20	48	7,302	7,246	252	441	4.9	4.8	5.1	5.1
F. graminearum	40	48	13,332	12,426	3,096	3,126	1.8	1.7	2.3	2.3
F. oxysporum	60	48	17,735	20,843	4,409	6,222	1.7	1.6	2.3	2.3
F. verticillioides	42	48	14,179	14,716	3,536	3,922	1.8	1.7	2.4	2.4
M. grisea	40	51	12,841	11,850	3,000	2,916	1.7	1.6	2.2	2.1
N. crassa	39	49	9,826	9,679	1,832	2,304	1.8	1.5	2.2	1.9
R. oryzae	40	36	17,467	16,477	3,413	3,962	2.3	3.0	3.8	3.5
S. pombe	12	36	5,055	4,913	2,764	2,616	0.9	1.0	2.0	2.2
S. sclerotiorum	39	51	14,522	11,119	3,278	2,490	1.8	1.8	2.3	2.4
S. nodorum	37	51	16,597	13,707	2,359	3,582	1.7	1.6	2.0	2.2

Table 5.2 Accuracy of prediction of gene structure elements (Sn/Sp). Data is provided for the algorithm with the original (GeneMark-ES) and with the enhanced intron submodel (GeneMark-ES-2). The Sn and Sp values were determined for the test sets of complete genes (test sets Type I, Table 3.2). Bold font shows the larger value out of the two adjacent ones. Differences in prediction accuracy are shown in the columns labeled δ.

		S. pombe			*C. immitis*			*F. verticillioides*			*M. grisea*			*C. cinereus*		
		Intron model			Intron model			Intron model			Intron model			Intron model		
		original	new	δ	original	new	δ	original	new	δ	original	new	δ	original	new	δ
	Sn	80.3	**88.2**	7.9	72.3	**82.8**	10.5	79.4	**85.6**	6.2	70.8	**89.2**	18.4	81.5	**85.0**	3.5
Internal exon	Sp	87.0	**89.6**	2.6	87.8	**93.0**	5.2	88.7	**91.2**	2.5	84.0	**91.7**	7.7	87.9	**89.7**	1.8
	Sn	84.0	**91.0**	7.0	75.4	**84.1**	8.7	85.6	**90.7**	5.1	76.6	**89.3**	12.7	84.7	**86.8**	2.1
Intron	Sp	89.1	**92.7**	3.6	85.6	**91.3**	5.7	91.5	**94.3**	2.8	84.2	**90.5**	6.3	89.2	**90.3**	1.1
	Sn	89.1	**93.1**	4.0	81.9	**87.0**	5.1	89.4	**92.2**	2.8	85.0	**92.1**	7.1	88.5	**89.6**	1.1
Donor	Sp	94.8	**95.1**	0.3	93.6	**94.4**	0.8	95.9	**96.2**	0.3	**94.4**	93.9	-0.5	**93.8**	93.6	-0.2
	Sn	86.1	**92.8**	6.7	77.7	**86.8**	9.1	87.1	**91.9**	4.8	79.7	**93.2**	13.5	86.1	**87.8**	1.7
Acceptor	Sp	91.3	**94.6**	3.3	88.3	**94.7**	6.4	93.3	**96.2**	2.9	87.6	**94.6**	7.0	91.2	**92.1**	0.9
	Sn	82.4	**88.0**	5.6	71.4	**79.7**	8.3	81.2	**85.3**	4.1	76.5	**88.0**	11.5	78.7	**81.2**	2.5
Exon	Sp	85.8	**89.2**	3.4	78.2	**84.6**	6.4	85.0	**87.9**	2.9	82.0	**89.1**	7.1	82.6	**84.3**	1.7
	Sn	85.8	**88.2**	2.4	75.9	**78.7**	2.8	81.0	**81.7**	0.7	84.6	**88.2**	3.6	**72.5**	**72.5**	0.0
Initiation site	Sp	86.4	**88.5**	2.1	76.8	**78.7**	1.9	81.3	**81.7**	0.4	86.1	**89.2**	3.1	**73.3**	72.9	-0.4
	Sn	92.7	**94.2**	1.5	82.4	**86.1**	3.7	92.4	**94.8**	2.4	79.9	**89.3**	9.4	80.8	**83.8**	3.0
Termination site	Sp	92.6	**94.2**	1.6	82.4	**87.1**	4.7	92.4	**95.4**	3.0	79.9	**89.3**	9.4	81.8	**84.8**	3.0
	Sn	98.1	**98.6**	0.5	94.7	**96.1**	1.4	97.9	**98.8**	0.9	95.8	**98.2**	2.4	**95.8**	95.3	-0.5
Nucleotide	Sp	99.4	**99.6**	0.2	95.3	**96.5**	1.2	96.5	**97.1**	0.6	93.5	**95.8**	2.3	94.6	**95.1**	0.5

Table 5.3 Accuracy of prediction of gene structure elements (Sn/Sp). Data is provided for the algorithm with the original (GeneMark-ES) and with the enhanced intron submodel (GeneMark-ES-2). The Sn and Sp values were determined for the test sets of incomplete genes (test sets Type II). Bold font shows the higher value out of the two adjacent ones. Differences in prediction accuracy are shown in the column labeled δ.

		A. nidulans			*A. niger*			*A. terreus*			*B. cinerea*			*C. neoformans*			*F. graminearum*		
		intron submodel			intron submodel			intron submodel			intron submodel			intron submodel			intron submodel		
		original	new	δ	original	new	δ	original	new	δ	original	new	δ	original	new	δ	original	new	δ
Internal exon	Sn	77.3	**87.4**	10.1	85.0	**91.5**	6.5	85.5	**91.6**	6.1	79.5	**87.9**	8.4	85.7	**92.3**	6.6	88.6	**92.6**	4.0
	Sp	90.5	**93.1**	2.6	91.4	**96.3**	4.9	90.9	**94.8**	3.9	91.4	**96.5**	5.1	91.1	**95.1**	4.0	93.6	**95.9**	2.3
Intron	Sn	81.1	**89.0**	7.9	86.2	**91.7**	5.5	88.2	**92.7**	4.5	84.7	**89.8**	5.1	86.8	**92.4**	5.6	90.5	**93.5**	3.0
	Sp	93.1	**96.4**	3.3	93.4	**96.8**	3.4	94.5	**97.4**	2.9	94.1	**96.7**	2.6	93.0	**96.0**	3.0	96.0	**97.5**	1.5
Donor	Sn	84.9	**90.5**	5.6	90.1	**92.9**	2.8	90.7	**93.7**	3.0	88.4	**91.1**	2.7	91.3	**94.6**	3.3	93.2	**94.7**	1.5
	Sp	95.6	**96.8**	1.2	96.2	**97.3**	1.1	96.1	**97.7**	1.6	97.1	**97.3**	0.2	96.4	**97.4**	1.0	97.8	**97.9**	0.1
Acceptor	Sn	83.8	**91.4**	7.6	89.3	**94.2**	4.9	90.2	**94.4**	4.2	87.0	**92.4**	5.4	88.7	**94.0**	5.3	92.0	**95.6**	3.6
	Sp	94.5	**97.5**	3.0	95.1	**98.5**	3.4	95.2	**97.9**	2.7	95.4	**98.6**	3.2	94.4	**97.2**	2.8	96.5	**98.5**	2.0

		R. oryzae			*F. oxysporum*			*N. crassa*			*S. sclerotiorum*			*S. nodorum*		
		intron submodel			intron submodel			intron submodel			intron submodel			intron submodel		
		original	new	δ	original	new	δ	original	new	δ	original	new	δ	original	new	δ
Internal exon	Sn	88.7	**88.8**	0.1	84.1	**92.5**	8.4	81.2	**85.2**	4.0	82.6	**90.2**	7.6	82.8	**88.5**	5.7
	Sp	94.3	**94.7**	0.4	87.8	**90.6**	2.8	92.0	**95.6**	3.6	91.3	**94.1**	2.8	90.7	**94.8**	4.1
Intron	Sn	88.8	**88.9**	0.1	86.7	**91.3**	4.6	85.9	**88.6**	2.7	86.3	**91.3**	5.0	87.3	**90.8**	3.5
	Sp	95.9	**95.9**	0.0	94.0	**94.8**	0.8	94.8	**97.0**	2.2	94.7	**96.4**	1.7	94.9	**97.2**	2.3
Donor	Sn	91.3	**91.4**	0.1	89.3	**93.4**	4.1	88.4	**89.6**	1.2	90.5	**93.5**	3.0	90.4	**92.4**	2.0
	Sp	97.0	**97.2**	0.2	95.0	**95.5**	0.5	96.7	**97.6**	0.9	**97.5**	97.4	-0.1	96.6	**97.6**	1.0
Acceptor	Sn	90.3	**90.4**	0.1	89.3	**94.3**	5.0	88.9	**91.3**	2.4	88.2	**93.8**	5.6	89.8	**93.3**	3.5
	Sp	96.7	**96.8**	0.1	95.4	**96.6**	1.2	96.8	**98.7**	1.9	95.5	**97.8**	2.3	96.1	**98.2**	2.1

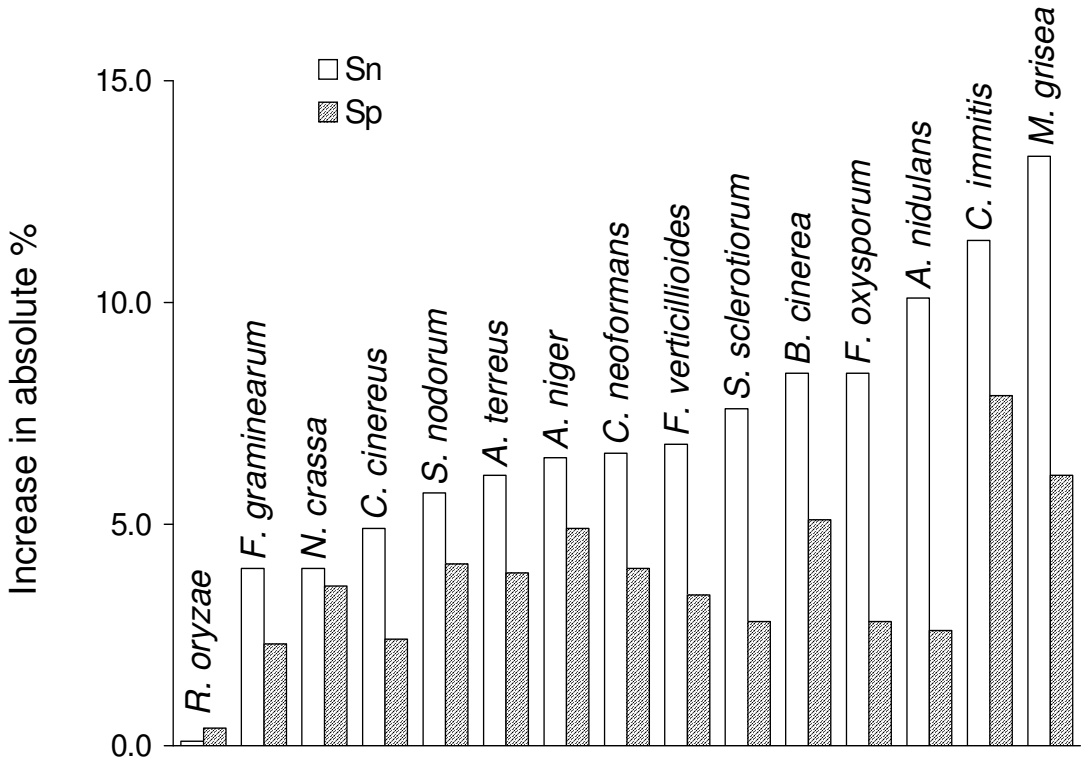


Figure 5.5 The increase in accuracy values of internal exon prediction. GeneMark-ES-2 shows significant improvement in accuracy of internal exon prediction. Marginal improvement for *R. oryzae* reflects the acceptor site upstream composition of this species.

Most of the sixteen species analyzed typically exhibit well conserved BP site (Figure 5.6) and acceptor that has rather weak signal (Figure 5.7). This is not, however, the case with *R. oryzae* and in fact, the opposite is observed; the acceptor site contains long nearly 20nt poly-Y tail and very weak BP site (Figure 5.6 and 5.7). Similarly, introns in *Phycomyces blakesleeanus* another representative of Zygomycota also contain weakly conserved BP sites and long poly-Y tails upstream to acceptor sites (Bruce M., Lomsadze A. and Borodovsky M., unpublished).

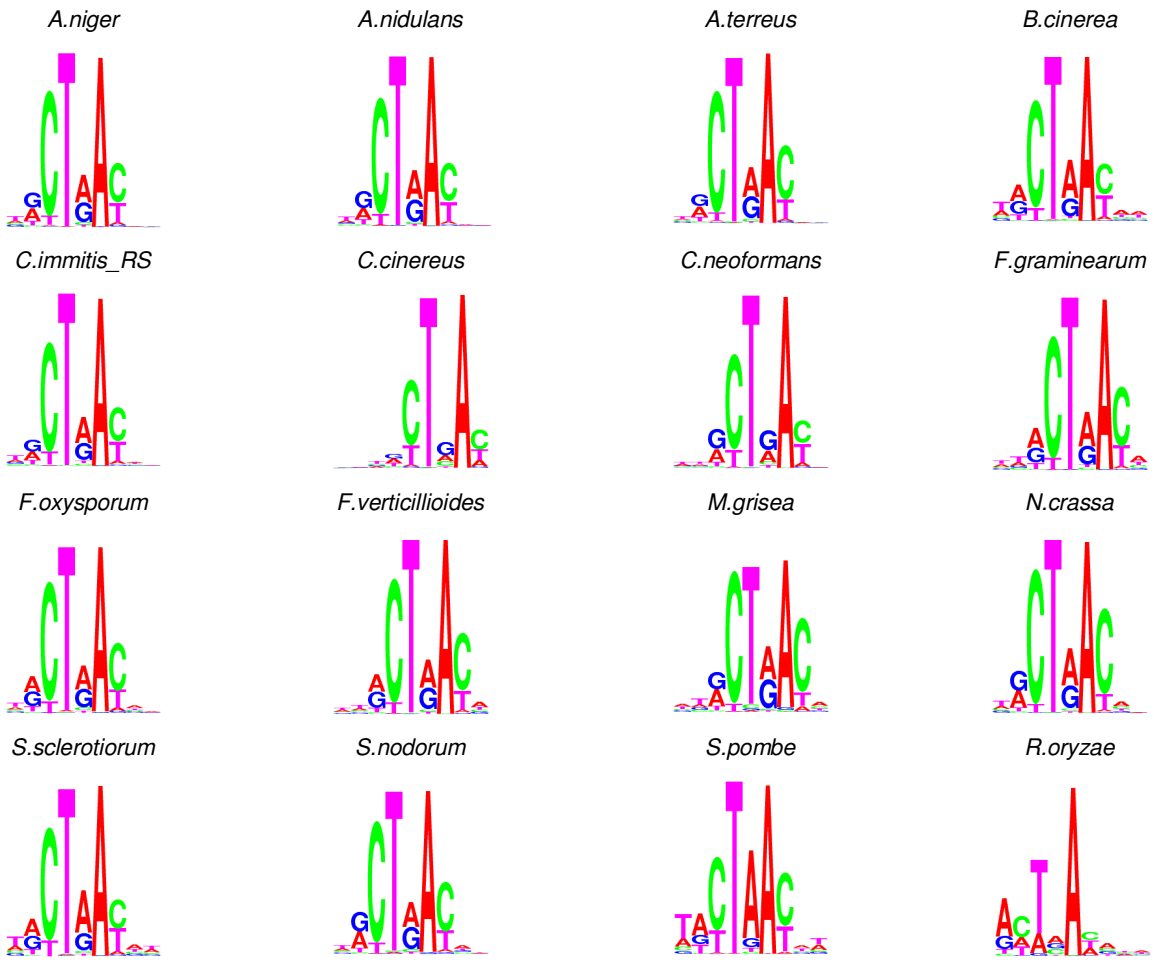


Figure 5.6 Zero order branch point model logos for sixteen fungal genomes determined by Gibbs sampling alignment of introns predicted at the final step of the algorithm.

Figure 5.7 Logos (from left) for *N. crassa* (Ascomycota), *C. neoformans* (Basidiomycota) and *R. oryzae* (Zygomycota). Introns which lack poly-Y tail possess conserved BP site. The models are derived from a set of acceptors predicted at the algorithm convergence.

In such situations the algorithm most frequently uses the lower path of the intron model which is in essence the same as running the original self-training. Table 5.4 shows the frequency of the upper and the lower paths chosen by the algorithm in iterations. While for the most species the best sequence of the hidden states follows the upper path with frequency of more than 90% for most of the species (except for *M. grisea* 77%) for *R. oryzae* this number is as low as 23%. Clear distinction splicing properties exhibited in *R. oryzae* and possibly other Zygomycota suggests that (i) the role of the BP model in intron prediction and in the splicing mechanism is relatively small and (ii) these species evolved under different selective pressure.

The spacer length distribution obtained at the algorithm convergence point shows a skewed bell shape (Figure 5.8) which allows better discrimination between true BP and other high scoring motifs in the acceptor's upstream region.

Table 5.4 Upper and lower path counts of the Viterbi parse shows that for the most species the best path of the hidden states follows through the upper path. *M. grisea* and *R. oryzae* are exceptions where the upper path is chosen in 77% and 23% of the time, respectively.

	Iteration index 6			Iteration index 7		
	Instances of upper path transition	Instances of lower path transition	% of upper path transition	Instances of upper path transition	Instances of lower path transition	% of upper path transition
A. nidulans	17,336	699	96.1	17,553	759	95.9
A. niger	22,368	1,069	95.4	22,592	1,172	95.1
A. terreus	21,263	1,150	94.9	21,605	1,132	95.0
B. cinerea	13,454	1,028	92.9	13,569	1,077	92.6
C. immitis	13,073	762	94.5	13,382	783	94.5
C. cinereus	50,321	3,079	94.2	53,454	3,313	94.2
C. neoformans	29,124	1,092	96.4	30,085	1,200	96.2
F. graminearum	5,982	261	95.8	6,025	264	95.8
F. oxysporum	31,355	2,127	93.6	31,408	2,246	93.3
F. verticillioides	24,296	1,034	95.9	24,359	1,149	95.5
M. grisea	12,568	5,576	69.3	14,109	4,120	77.4
N. crassa	12,945	1,264	91.1	12,880	1,358	90.5
R. oryzae	15,152	28,842	34.4	10,238	34,546	22.9
S. pombe	4,712	226	95.4	4,718	227	95.4

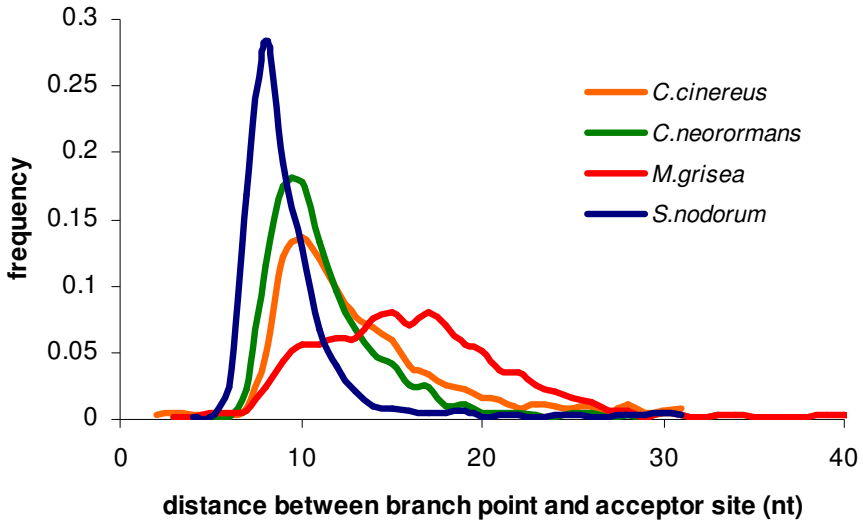


Figure 5.8 Length distribution of the sequences between branch point and acceptor site determined for four fungal species at the final iteration of the algorithm.

5.3.2. Accuracy of Gene Prediction on *S. pombe* Artificial Chromosomes

The ideal test set is described by curated segments (records) of a chromosome in which adjacent genes are separated by the intergenic region. The careful annotation of intergenic regions, as important as that of the genes, is necessary to accurately reflect the Sp values of the prediction program. Given the strict requirements in the procedure of test set derivation as described in Chapter 3, it is a challenging task to select a large set of records containing several genes in a row. Hence, most of the records in the test sets of Type I are represented by isolated genes. A test set of this nature is not suitable for identification of the errors associated with gene merging and gene splitting. For this reason, a sets of "artificial chromosomes" from 1,277 verified complete genes of *S. pombe* are constructed. Genes in the artificial chromosomes are placed in the 5' to 3' direction and connected by random sequences which compositionally reflect *S. pombe*'s intergenic region. Each one of these chromosomes is characterized by its set of genes and intergenic region for which the length is fixed and ranges from 50nt to 6,000nt (see Section 3.4.2 for details). A similar approach was applied by Pavy et al. (Pavy 1999) in test set preparation of their algorithm testing procedure.

The results shown in Figure 5.9 indicate that the new intron submodel is less prone to gene merging than the original intron model as tested on the set of artificial chromosomes. More than 50% of the intergenic length distribution derived from the whole *S. pombe* genome falls within the range of 750 to 6,000 (Figure 5.10). When taking into account also the fact that in reality not all of the genes are positioned on the direct strand the total number of gene merging for GeneMark-ES-2 is estimated to be 10 per 500 genes.

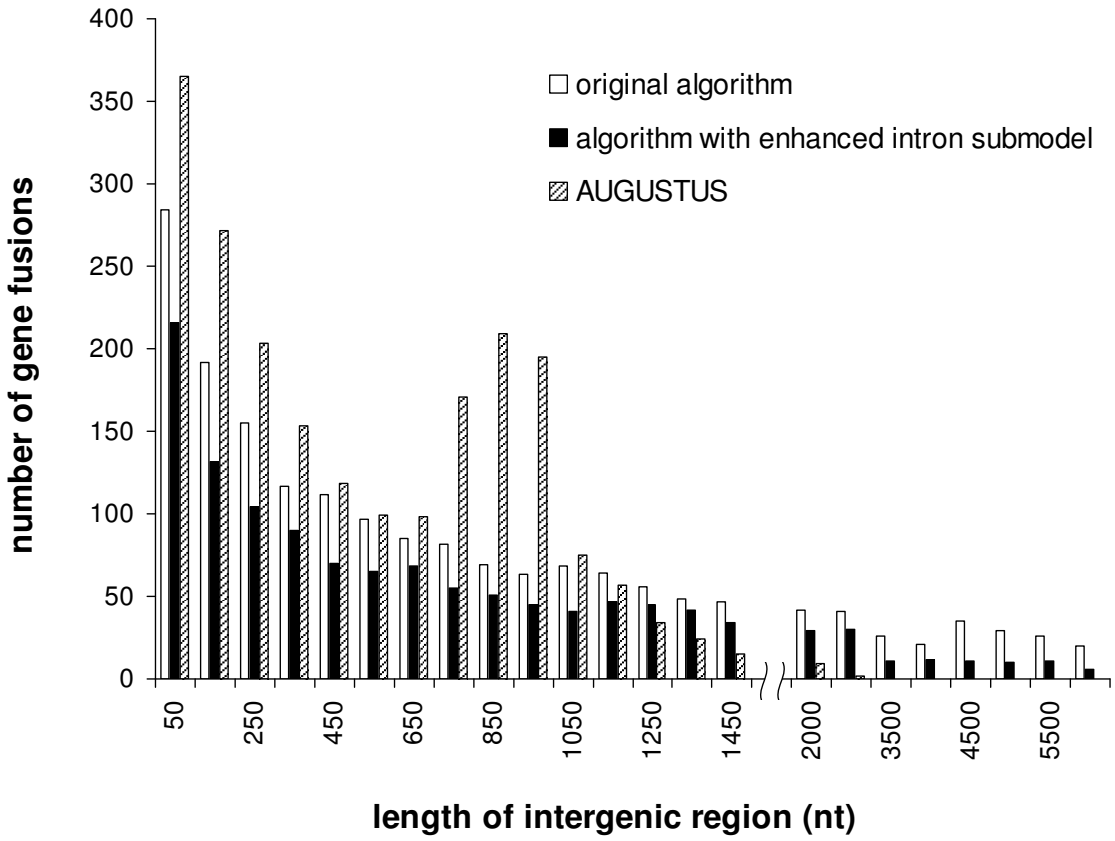


Figure 5.9 The number of gene fusions that occurred as a result of predictions in the set of artificial chromosomes. Each bin on the *x-axis* represents a chromosome where genes are connected by a random sequence which compositionally reflects the intergenic region of *S. pombe*.

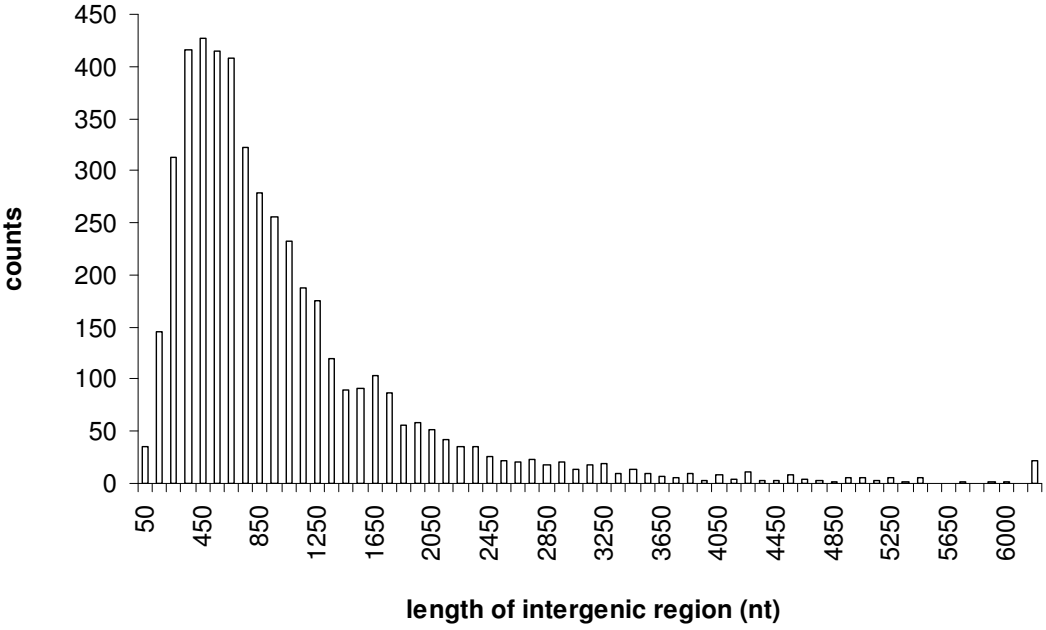


Figure 5.10 The length distribution of intergenic regions in the *S. pombe* genome (as annotated).

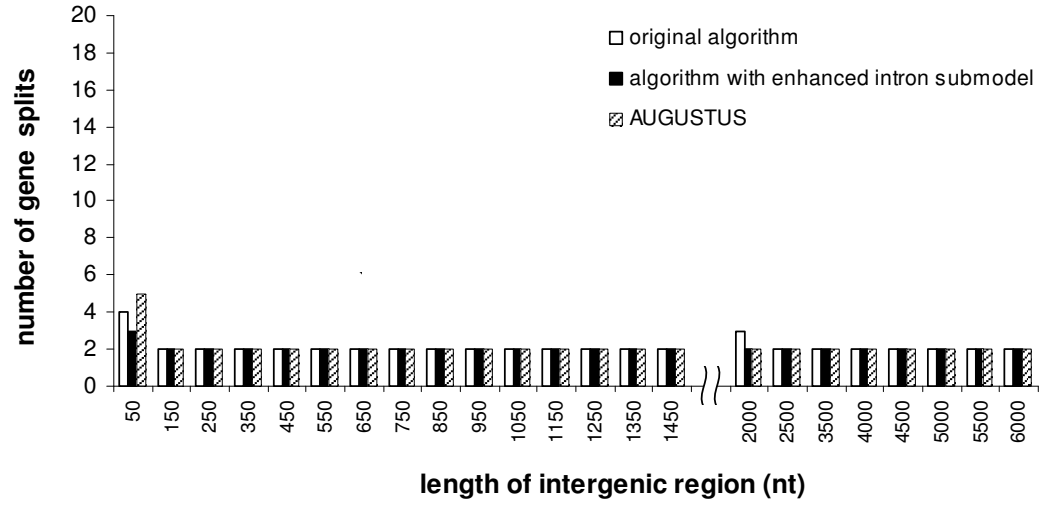


Figure 5.11 The number of gene splits that occurred as a result of predictions in the set of artificial chromosomes. Each bin on the *x-axis* represents a chromosome where genes are connected by a random sequence which compositionally reflects the intergenic region of *S. pombe*.

With regards to gene splitting unsupervised approaches (original intron model and new intron submodel) show no significant difference (Figure 5.11). In 1,277 genes GeneMark-ES-2 splits on average total of 2-3 times. The same is true for original the algorithm.

Interestingly, in this case a total of four splits is observed for artificial chromosome with intergenic lengths equal to 50 nt, confirming its weaker discrimination power between the intergenic region and true introns. The new algorithm performs better than the original algorithm in terms of predicting genes in intergenic regions (Figure 5.12). Given the actual distribution of length of *S. pombe* intergenic regions (Figure 5.10) the number of predicted false genes is estimated at 3 per 1,000 intergenic regions.

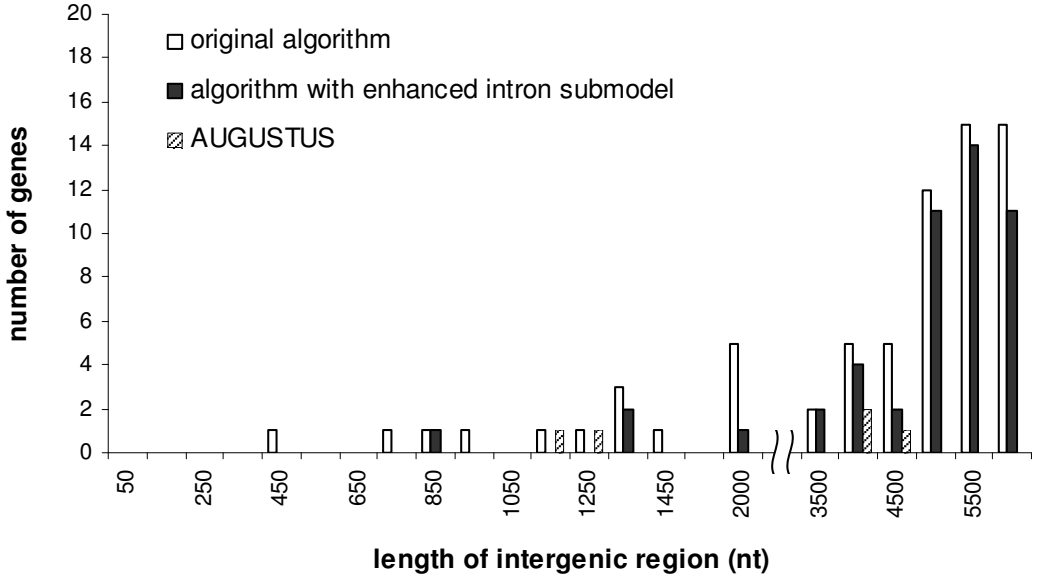


Figure 5.12 The number of genes that are predicted in intergenic (random) regions of *S. pombe* artificial chromosomes. Each bin on the *x-axis* represents a chromosome where genes are connected by a random sequence which compositionally reflects the intergenic region *S. pombe*.

The rate of exact gene prediction is also addressed. Figure 5.13 shows the dependence of the number of exactly predicted genes in artificial chromosomes on the length of intergenic region. Overall GeneMark-ES-2 shows better performance over the model obtained by original self-training. For short intergenic lengths (below 750 nt) GeneMark-ES shows a decrease in accuracy values. In contrast, the new algorithm is stable and on average exactly predicts 900 genes out of 1,277 (Figure 5.13); considering artificial chromosomes of entire range of intergenic lengths the original algorithm overall makes an average of 800 exact gene predictions.

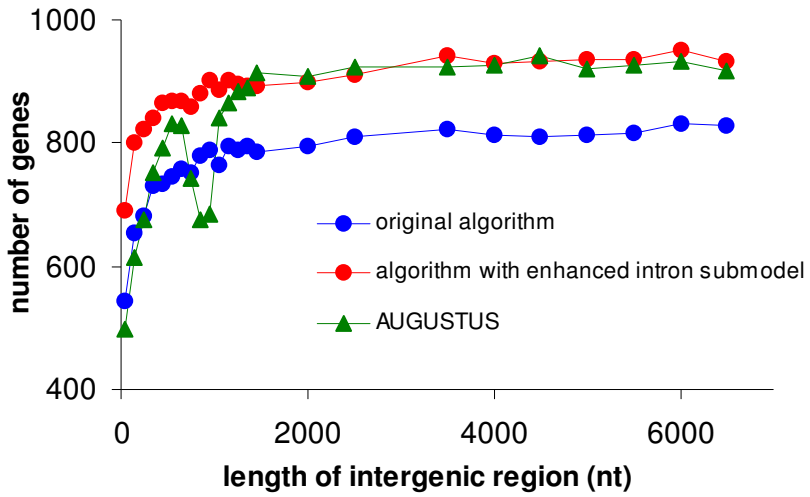


Figure 5.13 The number of exactly predicted genes in *S. pombe* artificial chromosomes vs. the length of intergenic region.

5.3.3. Dynamics of Convergence in Iterations

The accuracy values shown in Tables 5.2, 5.3, and 5.5 reflect GeneMark-ES-2 performance with the model obtained at the algorithm convergence. It is important to trace the values characterizing prediction accuracy in each iteration; this, for example,

may reveal situations when the final model has not converged to the global maxima. Similar, to the approach taken in Chapter 4, the changes in the accuracy of gene prediction in iterations is calculated by using the statistical models derived at a particular iteration with respect to the test set Type I. The results in terms of Sn and Sp of exon-intron structures are shown in Figure 5.14. The relatively low accuracy in the first iteration (iteration index 1) is not surprising since here, as with GeneMark-ES the iterations start with weak initial model (see Section 4.3.1 for details). The parse of the initial models, however, provides a training set for re-estimation of parameters to be used in the next iteration. The rate of specificity gain continues to grow and at the 3rd iteration it reaches 60-80%. The sensitivity however shows growth up to 60%. At the end of the 4th iteration (iteration index 4) where the state durations and phase dependencies are accounted for the Sn values grow up to 80%. Further improvement continues as the new intron submodel path is activated.

The increase in Sp is an important indication for successful self-training. As the training set in iterations is enriched with true predictions the models become more sensitive in their prediction power and when the additional parameters are "freed" the statistical model is able to find the gene elements with about 80% accuracy.

5.3.4. Convergence

The average size of a fungal genome considered in this study is relatively small compared to that of genomes presented in Chapter 3 (e.g. *A. thaliana, D. melanogaster and C. intestinalis*). In addition the parameter space for GeneMark-ES-2 is more

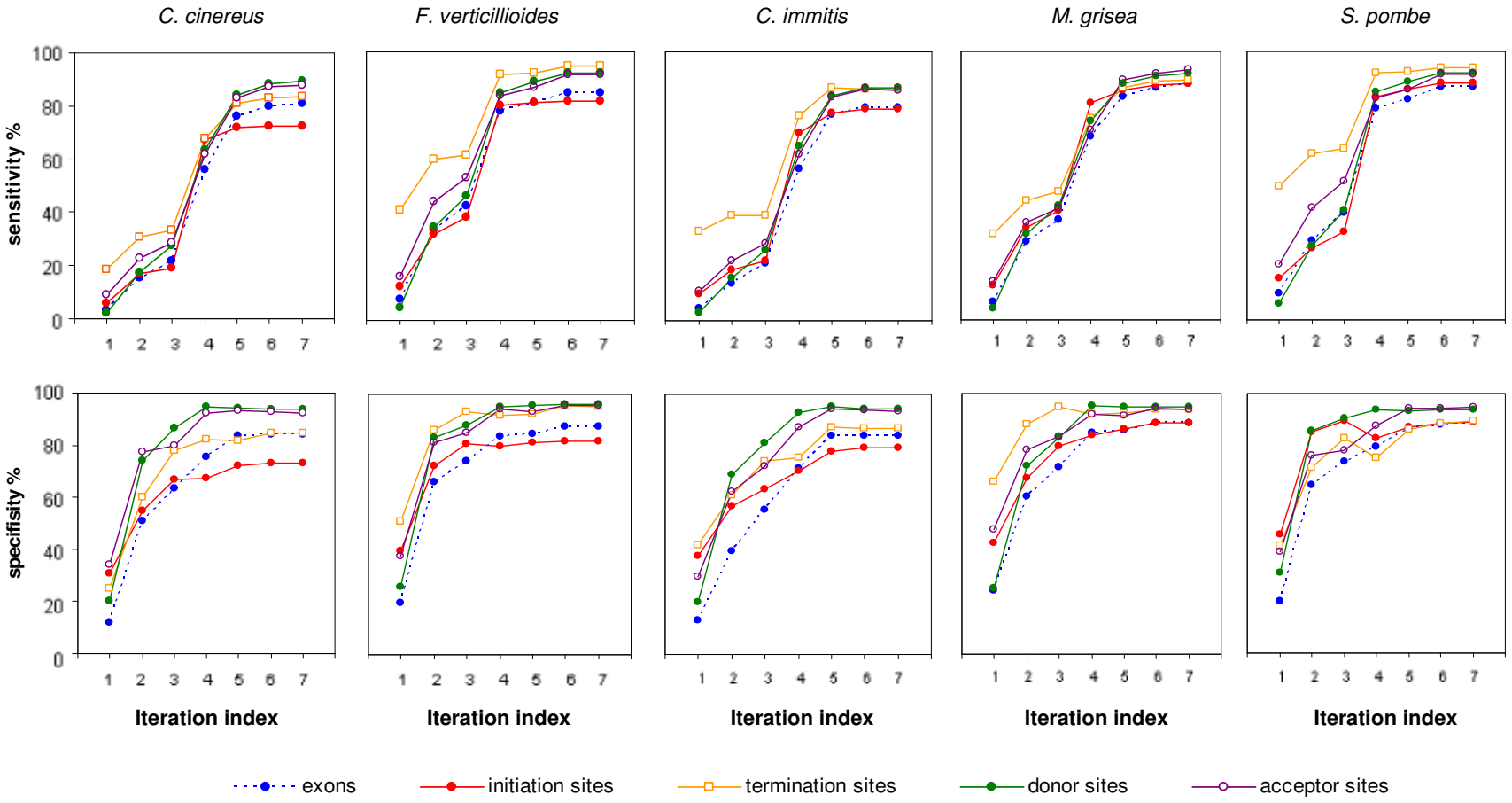


Figure 5.14 Changes of Sn and Sp of exon-intron structure prediction in iterations for five fungal species.

complex. Nevertheless the algorithm is stabilized at iteration index 7 for all fungi species (compared to 6 for species described in Chapter 3). Interestingly, at the convergence the two parses of the last iterations contained not only nearly identically labeled nucleotide sequences but also in 97%-99% matching positions of translation termination sites.

5.3.5. Intron Submodel Features at the Algorithm Convergence

Upon completion of iterations the signal site model parameters are compared to those derived from extrinsically determined introns which are based on EST to genomic sequence alignment. The number of introns determined by EST alignments vary from 1,152 (*Fusarium oxysporum*) to 7,812 (*C.cinereus*). These numbers are 3 to nearly 50 times larger for the set of introns that is obtained at the GeneMark-ES-2 convergence and applied to the whole genomic sequence. Table 5.5 shows the KL distances between a particular model that belongs to the lower path of GeneMark-ES-2 (Figure 5.3) and the background sequence or uniform distribution in case of the downstream spacer. The first order models are used in KL value calculations for both unsupervised training and alignment method. The difference for the majority of cases is not more than 7%. The branch point motif on the other hand, as a rule, carries an amount of information comparable to what is seen for donor sites.

For all species with the exception of *R. oryzae* the branch point exhibits a motif stronger than the acceptor site signal which agrees with the accuracy results.

In the predicted BP sites of all of the sixteen fungal species except for *R. orizae*, the consensus sequence of the positional nucleotide frequencies is CTNAC (Figure 5.6). The estimated BP frequencies related to the canonical “A” in the BP position vary from 97%

Table 5.5 Relative entropies of the first order models of donor, branch point and acceptor sites as well as the length distributions of the downstream spacers derived from the sets of intron determined by (i) the self-training algorithm and (ii) EST to genome alignment. Differences between the values derived by different methods are shown in columns labeled δ.

Species	**donor**			**branch point**			**acceptor**			**spacer**		
	self-training	alignment	δ	self-training	alignment	δ	self-training	alignment	δ	self-training	alignment	δ
A. niger	8.0	7.8	0.2	7.3	7.6	-0.3	5.1	5.0	0.1	2.1	1.7	0.4
A.nidulans	7.7	7.6	0.1	7.3	7.4	-0.1	5.0	5.0	0.0	2.0	1.8	0.2
A. terreus	7.9	7.7	0.2	7.5	8.0	-0.5	5.1	5.1	0.0	2.1	2.1	0.0
B. cinerea	7.9	8.2	-0.3	7.4	8.2	-0.8	5.0	5.1	-0.1	2.2	2.4	-0.2
C. immitis	7.8	7.4	0.4	7.2	7.0	0.2	5.3	5.0	0.3	1.9	1.4	0.5
C. cinereus	7.9	7.8	0.1	5.7	6.3	-0.6	5.3	5.3	0.0	1.1	0.9	0.2
C. neoformans	8.5	7.1	1.4	6.7	5.9	0.8	5.1	5.1	0.0	1.8	1.8	0.0
F. graminearum	8.4	8.6	-0.2	7.6	8.3	-0.7	5.0	5.0	0.0	2.3	2.5	-0.2
F. oxysporum	7.5	8.7	-1.2	7.2	8.1	-0.9	4.8	5.6	-0.8	1.7	2.3	-0.6
F. verticillioides	8.2	8.3	-0.1	7.5	7.8	-0.3	4.9	5.2	-0.3	2.0	2.3	-0.3
M. grisea	7.9	8.5	-0.6	7.5	8.2	-0.7	4.9	5.3	-0.4	1.1	1.6	-0.5
N. crassa	8.7	8.5	0.2	8.3	8.2	0.1	5.1	5.3	-0.2	2.4	1.7	0.7
R. oryzae	7.1	5.4	1.7	4.0	4.1	-0.1	5.1	6.4	-1.3	0.3	0.8	-0.5
S. sclerotiorum	7.8	8.2	-0.4	7.3	7.8	-0.5	5.0	5.2	-0.2	2.0	2.7	-0.7
S. pombe	8.6	9.2	-0.6	7.6	7.8	-0.2	5.4	7.2	-1.8	1.8	1.8	0.0
S. nodorum	7.6	8.5	-0.9	7.2	7.8	-0.6	4.8	5.3	-0.5	2.1	2.2	-0.1

(*R. oryzae*) to 99% (*C. cinereus*). The most frequent length of the downstream spacer is in the range of 8-18 nt. The highest localization is observed for *S. nodorum* (Figure 5.8) which exhibits one of the highest spacer KL values (Table 5.5).

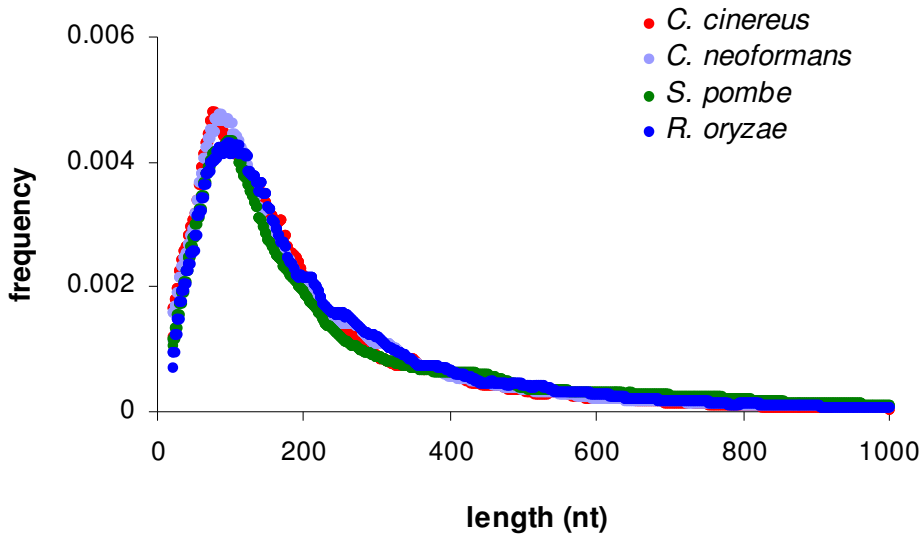
The value of the relative entropy for the spacer length distributions, with regard to the uniform distribution, shows the degree of "compactness" for the BP site localization. This value along with KL distance for BP site motif can be used to infer about (i) evolutionary conservation of BP motifs in given species (*R. orizae* for example) and (ii) possible problems within self-training procedure occurred due to false predictions.

5.3.6. State Durations

While the length distributions of exons obtained at the algorithm convergence step in all sixteen fungal genomes (Figure 5.15) are localized at 100 nt two types of distribution in terms of shape are noticed. First, with a weak localization (Figure 5.15a) is observed for most of genomes in the phylum Ascomycota (Figure 5.15b). Second, a more localized exon length distribution is observed for *C. cinereus, C. neoformans,* (Basidiomycota*), S. pombe* (Ascomycota) and *R. oryzae* (Zygomycota).

The Figure 5.16 shows three distinct shapes for intron length distributions with low (Figure 5.16a), intermediate (Figure 5.16b) and high (Figure 5.16c) localizations. Not surprisingly, species of the same genus were clustered together. The differences are seen at the level of *order*. For instance, while *Fusarium* species, representatives of *Sordariomycetes incertaesedis*, exhibit a highly localized intron length distribution (Figure 5.16a) *M.grisea* and *N.crassa* that belong to the same *order* have introns that have the lowest localization (Figure 5.15c).

a)



b)

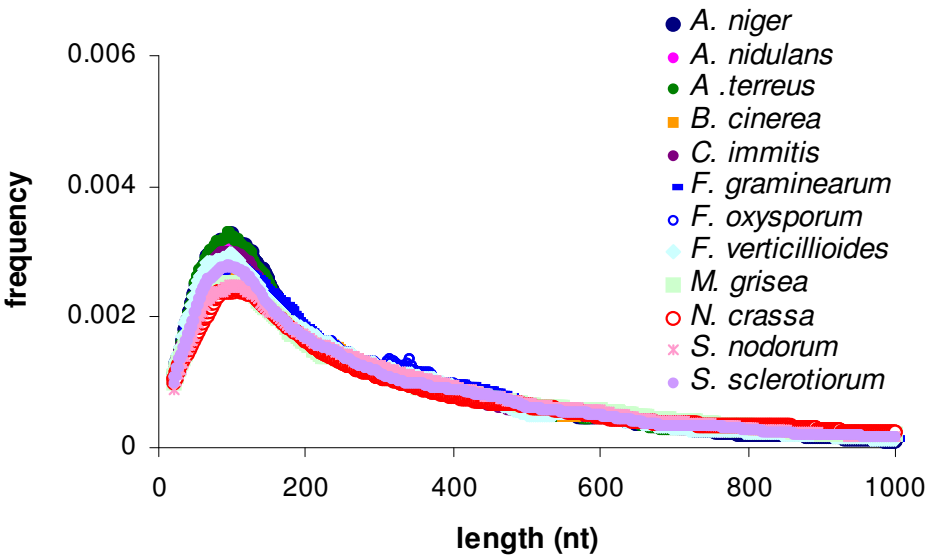


Figure 5.15 The exon length distributions as determined for sixteen fungal genomes at algorithm convergence.

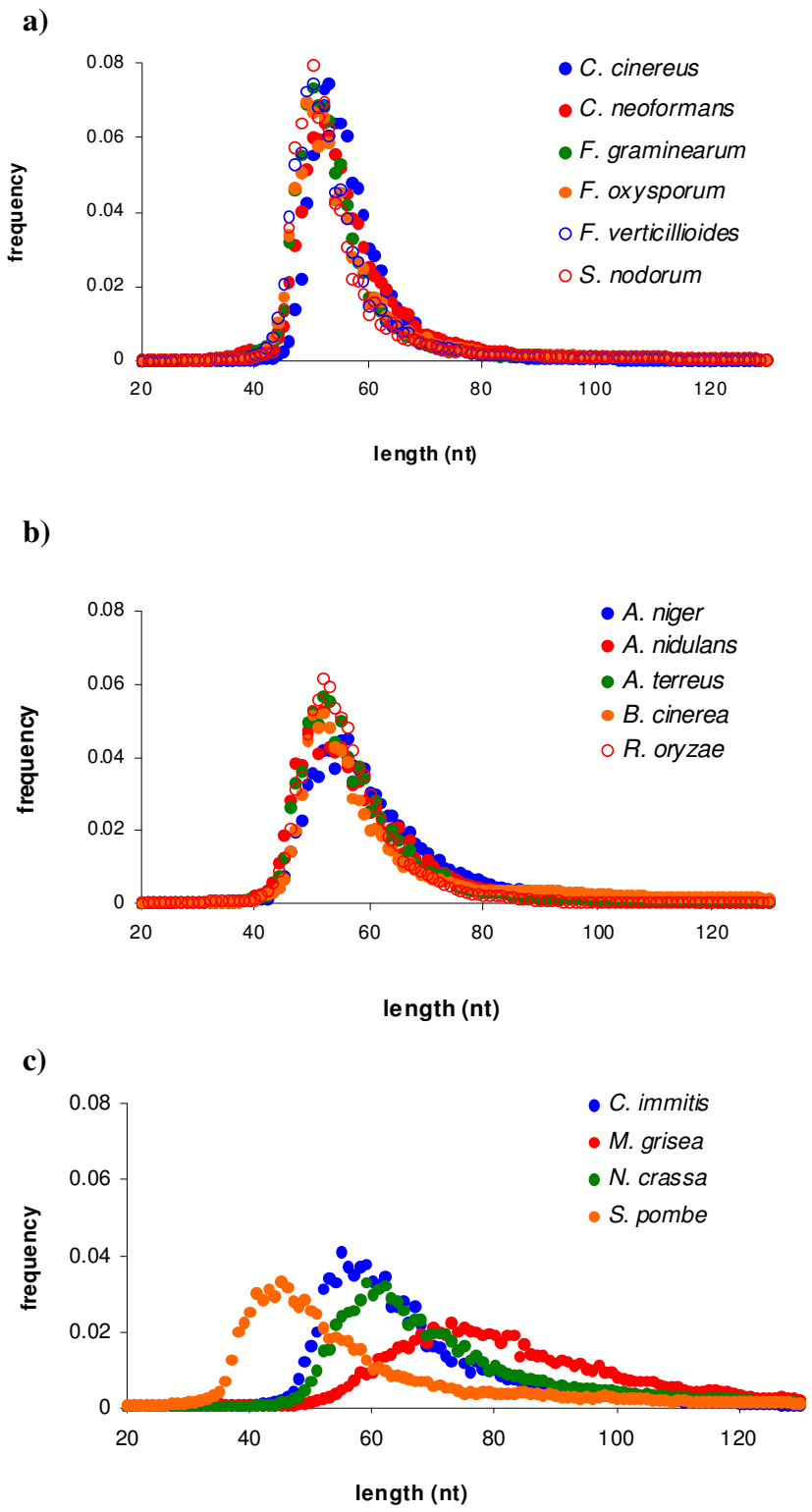


Figure 5.16 The intron length distributions as determined for sixteen fungal genomes at the algorithm convergence.

5.3.7. Inhomogeneous G+C content

The G+C distribution of the genomic sequences for the most of the fungal species presented in this chapter is observed to be homogeneous. Figure 5.8 shows the G+C histogram for four of these species. Only the genomic sequence of *C. immitis* among all has shown compositional inhomogeneity (Figure 5.17). GeneMark-ES-2 is applied twice: (i) on the whole genomic sequence and (ii) on the high G+C cluster. The accuracy results determined from the test set Type I show that GeneMark-ES-2 run with input (ii) produces about 1%, increase in average sensitivity and specificity of splice sites prediction in comparison with use of input (i). Furthermore, what is more important, the model obtained with input (ii) has a smaller rate of false positives in a randomly generated sequence reflecting *C. immitis* non-coding sequence. Therefore, the results of GeneMark-ES-2 with the clustered input are reported.

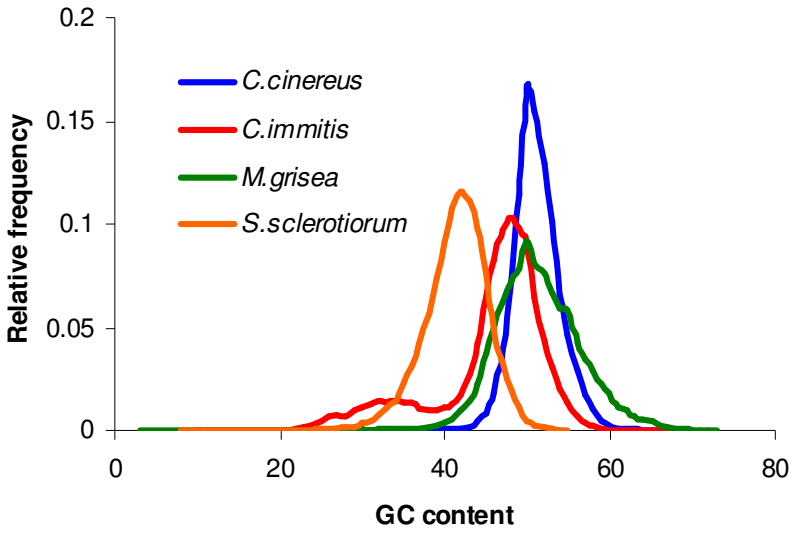


Figure 5.17 G+C histogram of genomic DNA determined for four fungal genomes. Histogram is calculated for 1 kb long non-overlapping fragments.

5.3.8. Comparison with Other Gene Prediction Programs

By definition the model parameters of supervised gene finders are based on the training set of validated genes. In order to assess the accuracy of such an approach the training set should not contain any genes from the test set. Frequently, the *n*-fold cross validation is applied in order to evaluate the performance of the predictor, (Korf 2004). For self-training approaches the sequences that contain test set data can be excluded from the input set.

However, in implementation of the unsupervised training, the knowledge from the set of annotated genomic data is not utilized since in such an approach the algorithm uses the set of unlabeled genomic sequences.

Re-training of supervised algorithms can be a cumbersome task for different reasons which include the availability, flexibility of the algorithm as well as well training manuals. The comparisons below are based on the basis of availability of programs and literature where the authors site the accuracies of a particular method.

The GipsyGene program is one of a few designed specifically for fungal genomes (Neverov 2003). The program is trained for *A. nidulans* and *N. crassa*. Given the difference in the test sets the practical comparison would be on the level of sensitivity values. The integration of the BP model in GipsyGene program resulted in increase of internal exon prediction sensitivity from 69-75% and 75-80% for *N. crassa* and *A. nidulans* respectively. The GeneMrak-ES-2 Sn results in the same category are 81.2% (*N. crassa*) and 77.3% (*A. nidulans*) for the original intron model and 85.2% (*N. crassa*) and 87.4% (*A. nidulans*) for the enhanced intron model (Table 5.3). Although the integration of the BP into GipsyGene leads to a significant increase in Sn values the

overall accuracy is still relatively low. These results are probably due to the size of the training set employed to derive the model parameters for GipsyGene. Sets of 193 and 99 genes used for *A. nidulans* and *N. crassa*, respectively, are rather small for accurate model parameterization; thus underlining the efficiency of unsupervised training in applications to genomes with small set of verified genes.

AUGUSTUS (Stanke and Waack 2003), also a supervised gene finding program (see Chapter 2 for details) which can be downloaded and run locally is used for comparison purposes. AUGUSTUS' performance is evaluated on test set Type I (Chapter 3). Table 5.6 shows that GeneMark-ES-2 outperforms AUGUSTUS in terms of average accuracy (Sn+Sp)/2 in all eight categories for *F. verticulloides*, *M. grisea* and *S. pombe* and in seven out of eight categories for *C. immitis* and *C. cinerius.*

Due to insufficient numbers of validated genes in the training set AUGUSTUS does not provide statistical models for *F. verticulloides*. For this reason the models from its close relative *F. graminearum* are used. The smallest gap in prediction accuracy between GeneMark-ES-2 and AUGUSTUS is expressed for *S. pombe*. Several reasons are likely to cause such a result. *S. pombe* is one of the well-studied fungal species. The number of experimentally known genes in this case is large. In fact, since the completion of the *S. pombe* sequencing project (Wood Gwilliam et al. 2002) its annotation is significantly improved over the years. Hence, *S. pombe* is a good candidate for supervised training. Another reason is that the test set contains about a half of all *S. pombe* multiple exon genes which greatly increases the chance of overlap between the test and the training sets. Finally, the size of this species, 12 Mb makes it the closest to

Table 5.6 Comparison of the performances of the GeneMark-ES-2 program and the AUGUSTUS program. Values of Sn and Sp were determined for the test sets of complete genes (test sets Type I, Table S4). For gene prediction in *F. verticillioides* the AUGUSTUS program uses model parameters derived in supervised mode for the *F. graminearum* genome. Bold font shows the larger value out of the two in corresponding category between AUGUSTUS and GeneMark-ES-2.

		S. pombe				*C. immitis*				*F. verticillioides*				*M. grisea*				*C. cinereus*			
		AUGUSTUS		GeneMark-ES-2		AUGUSTUS		GeneMark-ES-2		AUGUSTUS		GeneMark-ES-2		AUGUSTUS		GeneMark-ES-2		AUGUSTUS		GeneMark-ES-2	
Internal exon	Sn	87.7	88.1	**88.2**	**88.9**	**82.9**	83.7	82.8	**87.9**	82.9	81.4	**85.6**	**88.4**	80.9	83.7	**89.2**	**90.5**	82.2	84.4	**85.0**	**87.4**
	Sp	88.4		**89.6**		84.4		**93.0**		79.8		**91.2**		86.4		**91.7**		86.5		**89.7**	
Intron	Sn	90.2	91.8	**91.0**	**91.9**	82.9	86.2	**84.1**	**87.7**	85.2	87.5	**90.7**	**92.5**	79.3	83.5	**89.3**	**89.9**	82.7	85.9	**86.8**	**88.6**
	Sp	**93.3**		92.7		89.5		**91.3**		89.8		**94.3**		87.6		**90.5**		89.0		**90.3**	
Donor	Sn	92.6	93.9	**93.1**	**94.1**	86.3	88.6	**87.0**	**90.7**	90.8	90.8	**92.2**	**94.2**	85.2	89.0	**92.1**	**93.0**	86.4	89.0	**89.6**	**91.6**
	Sp	**95.1**		**95.1**		90.8		**94.4**		90.8		**96.2**		92.7		**93.9**		91.5		93.6	
Acceptor	Sn	91.2	92.8	**92.8**	**93.7**	85.1	88.5	**86.8**	**90.8**	87.3	90.0	**91.9**	**94.1**	82.1	87.6	**93.2**	**93.9**	84.9	88.3	**87.8**	**90.0**
	Sp	94.4		**94.6**		91.9		**94.7**		92.6		**96.2**		93.0		**94.6**		91.6		**92.1**	
Exon	Sn	85.9	87.4	**88.0**	**88.6**	76.9	80.0	**79.7**	**82.2**	76.7	78.5	**85.3**	**86.6**	78.7	83.6	**88.0**	**88.6**	78.3	81.3	**81.2**	**82.8**
	Sp	88.8		**89.2**		83.0		**84.6**		80.3		**87.9**		88.5		**89.1**		**84.3**		**84.3**	
Initiation site	Sn	83.9	87.0	**88.2**	**88.4**	74.3	**81.1**	**78.7**	78.7	70.6	77.3	**81.7**	**81.7**	76.9	85.9	**88.2**	**88.7**	70.7	**76.1**	**72.5**	72.7
	Sp	**90.1**		88.5		**87.9**		78.7		**84.0**		81.7		**94.9**		89.2		**81.4**		72.9	
Termination site	Sn	91.5	93.8	**94.2**	**94.2**	79.6	85.6	**86.1**	**86.6**	82.3	88.7	**94.8**	**95.1**	76.9	84.9	**89.3**	**89.3**	78.4	84.1	**83.2**	84.0
	Sp	**96.0**		94.2		**91.5**		87.1		95.1		**95.4**		**92.9**		89.3		**89.7**		84.8	
Nucleotide	Sn	96.2	97.9	**98.6**	**99.1**	90.4	93.3	**96.1**	**96.3**	95.9	96.0	**98.8**	**98.0**	87.8	92.1	**98.2**	**97.0**	90.9	92.9	**95.3**	**95.2**
	Sp	99.5		**99.6**		96.2		**96.5**		96.0		**97.1**		**96.4**		95.8		94.8		**95.1**	

the minimum size (10 Mb) required for unsupervised training (Chapter 4 and Lomsadze, Ter-Hovhannisyan et al. 2005). Nevertheless, the performance of the unsupervised algorithm with the new intron submodel is better in 13 out of 16 categories. AUGUSTUS is also used in analysis utilizing artificial chromosomes. The supervised prediction program demonstrates an error rate similar to the new algorithm in the category of gene splitting (Figure 5.11). While AUGUSTUS performs with a higher error rate associated with gene merging on the intergenic lengths below 1,150 nt it shows only marginal improvement over the new model for the longer intergenic lengths. A significant increase in the rate of gene merging is observed for chromosomes with intergenic length in the range of 750-950 nt. This possibly is related to the parameter settings or modeling approach used within the predictor, e.g. the conjunction point of state durations for short and long introns.

GeneMark-ES-2 demonstrates the best performance in terms of number of genes exactly predicted for the artificial chromosomes with shorter intergenic lengths. AUGUSTUS's sharp decrease in accuracy of exact gene prediction coincides with the spike observed for gene merging at 750-950 nt lengths (Figure 5.9). For the intergenic length above 1,450 nt, both unsupervised training with the new intron model and AUGUSTUS show similar accuracies (Figure 5.13). Again, it should be emphasized that AUGUSTUS results for *S. pombe* are on the highest level of its performance and are possibly overestimated.

5.3.9. Comparison with Annotation

The models obtained at the convergence step are used to run on whole genomic data which include the sequences that did not pass the pre-processing step described in Section 4.2.7. This run provides with predictions on whole genome level which then are translated into proteins. The latter is used in comparison with the annotation as described in Section 4.2.8.

GeneMark-ES-2 predictions are in agreement with existing annotations both in terms of the total number of genes and the number of introns per spliced gene (Table 5.1 and Figure 5.18) in genomes of *Aspergillus nidulans*, *Aspergillus terreus*, *C. cinerius*, *C. neoformans, F. verticillioides, N. crassa, S. pombe*. In other genomes i.e. *A. niger*, *B. cinerea, C. immitis, F. graminearum, F. oxysporum, M. grisea, R. orizae, S. nodorum*, *S. sclerotiorum* however the difference in total number of predicted and annotated genes is close to or exceeds 1,000 genes (Table 5.1).

Further analysis is carried out for these nine gnomes to address the significant differences between annotation and predictions. Novel genes in these genomes are determined as described in 4.4.8. The subset of annotated proteins that do not have similarity hit to predicted proteins are used in search against NR and CDD databases. Most of these proteins do not show statistically significant similarity to the entries in these databases (Table 5.7).

The fraction of GeneMark-ES-2 predictions however shows similarities to NR and CDD databases with a higher rate (Table 5.8). The biggest difference in the number of predicted genes is observed for *B. cinerea* and *F.* oxysporum (Table 5.1).

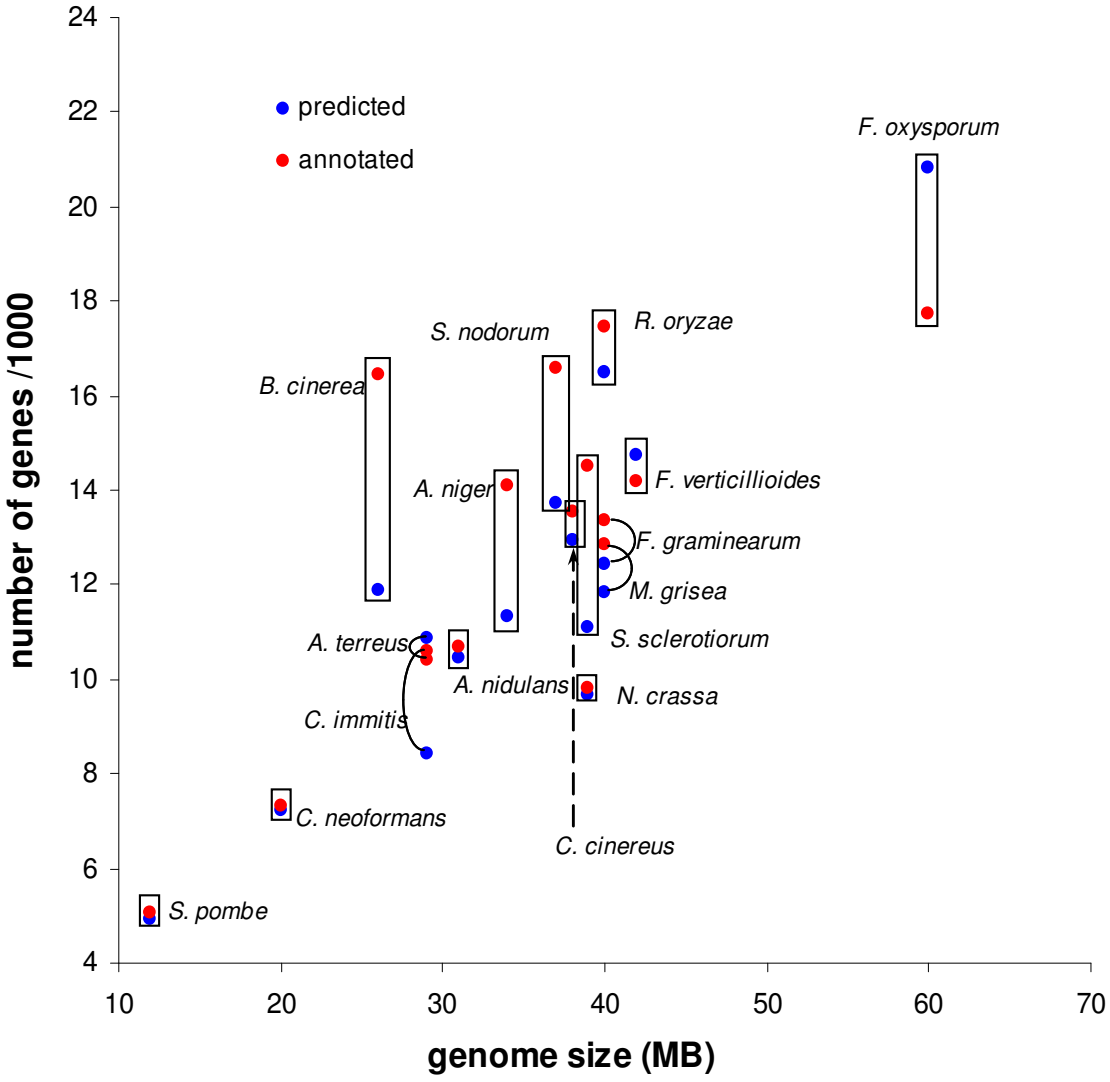


Figure 5.18 Genome size and the number of predicted and annotated genes in the sixteen fungal genomes.

Table 5.7 Analysis of annotated genes of nine fungal genomes.

Species	total number of annotated genes	number of protein products with no similarity to predicted proteins	proteins (from B) with similarity to proteins of other species in nr database	(%) of B	proteins (from B) with conserved domains (similarity to CDD)	(%) of B
	A	B	C	D	E	F
A. niger	14,101	2,851	107	3.75	26	0.91
B. cinerea	16,448	4,413	218	4.94	43	0.97
C. immitis	10,457	2,005	145	7.23	34	1.70
F. graminearrum	13,332	1,024	51	4.98	13	1.27
F. oxysporum	17,735	630	47	7.46	3	0.48
M. grisea	12,841	1,491	39	2.62	10	0.67
R. oryzae	17,467	1,086	17	1.57	10	0.92
S. sclerotiorum	14,522	3,655	226	6.18	29	0.79
S. nodorum	16,597	2,746	169	6.15	16	0.58

Table 5.8. Analysis of GeneMark-ES-2 predicted gene products of nine fungal genomes.

Species	total number of predicted genes	number of protein products with no similarity to annotated proteins	proteins (from B) with similarity to proteins of other species in nr database	(%) of B	proteins (from B) with conserved domains (similarity to CDD)	(%) of B
	A	B	C	D	E	F
A. niger	11,342	263	115	43.73	23	8.75
B. cinerea	11,890	529	153	28.92	40	7.56
C. immitis	8,435	311	94	30.23	29	9.32
F.graminearum	12,426	288	121	42.01	27	9.38
F. oxysporum	20,843	1,408	561	39.84	107	7.60
M. grisea	11,850	346	94	27.17	36	10.40
R. oryzae	16,477	446	108	24.22	71	15.92
S. sclerotiorum	11,119	342	112	32.75	21	6.14
S. nodorum	13,707	285	33	11.58	7	2.46

These two genomes represent the ends of the spectrum with *B. cinerea* being over-annotated (in comparison with GeneMark-ES-2 predictions) by 4,558 genes and *F. oxysporum* under- annotated by 3,108 genes.

The histograms of the protein lengths representing the subsets from annotation (*B. cinerea*) and GeneMark-ES-2 predictions (*F. oxysporum*) in which the proteins do not have similarity in the counterpart set are shown in Figure 5.19.

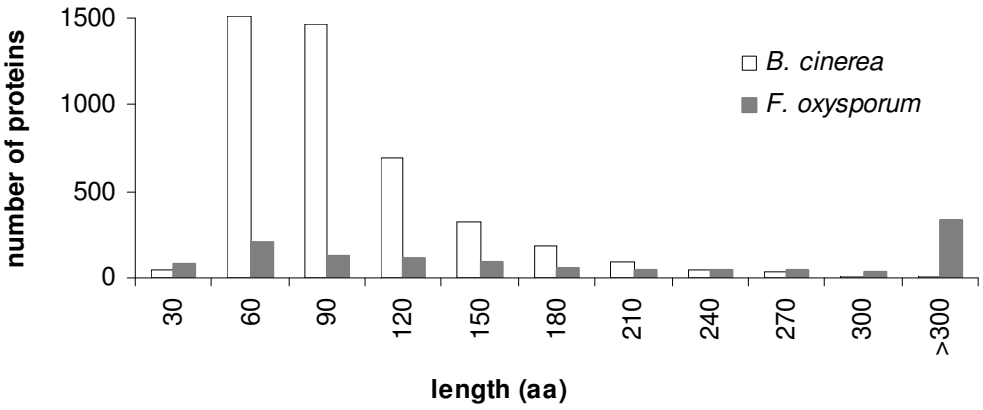


Figure 5.19 Length distributions of the annotated (*B. cinera*) and predicted (*F. oxysporum*) protein subsets. The proteins in each subset do not have similarity hit in predictions (*B. cinerea*) or annotation (*F. oxysporum*).

While the annotated subset exhibits excess in proteins with short lengths (*B. cinerea*) the predicted proteins (*F. oxysporum*) do not show a length bias. The short predictions correspond to the incomplete genes which are usually predicted at the end of contigs. Moreover, nearly 340 proteins in *F. oxysporum* have length of greater than 300aa. Many of the proteins in *F. oxysporum* subset show similarity to proteins

characterized as “hypothetical” or “conserved” which supports the predictions but does not provide information about the protein functionality.

5.4 Functionally Characterized New Genes

The complete list of the functional characterization of newly identified proteins in all sixteen fungal genomes is presented in Table A2 of the Appendix. The coordinates of the exon-intron genes in gff format is available at http://exon.gatech.edu/GeneMark/ gmhmm-es-2008. The novel genes are characterized by statistical significance (e-value which in this case ranges from e-113 to e-185). The most interesting findings are discussed in this section.

Proteins predicted in several fungal species

Presence of these proteins in different fungi increases the confidence of findings.

1. TOM7, mitochondrial outer membrane receptor is missing from annotation of five fungal species (*A. niger*, *B. cinerea*, *C. immitis*, *F.oxysporum* and *S. sclerotiorum*). Note that for *A. niger* and *B. cinerea* the number of predicted genes is considerably smaller than the number of annotated genes (Table 5.1).

2. Protein product of *A. niger* produced from new three-exon gene is homologous to Urm1. It exhibits nearly full length similarity to proteins in *A.oryzae, C.immitis, F.graminearum* and *B.cinerea. Urm1*, ubiquitin-like protein is involved both in regulation of invasive growth and budding in yeast as well as in regulation of nitrogen catabolite repression. The Urm1 protein is expected to be conserved since the urmilation process is

also detected in mammals. No paralogs of Urm1 gene are detected in the *A. niger* genome.

3. tRNA processing enzyme *RNAse P* found in *A. niger*, *C. immitis*, *F. graminearum* and *M. grisea*.

4. NADH-ubiquinone reductase protein associated with encephalomyopathies in humans is missing from annotation of *C. immitis*, *F. oxysporum* and *S. sclerotiorum.*

5. RNAse H responsible for digestion of RNA and RNA/DNA hybrid strands. While single exon gene encodes for this protein in *F. oxysporum* in *A. niger* this protein is produced by four-exon gene.

6. Isoprenylcystene carboxyl methyltransferase homologous to yeast Ste14 is identified in both *S. sclerotinia* and *B.cinerea*.

7. Skp1 protein, encoded by new two-exon gene in *A. niger* genome is a component of SCF complex involved in cyclin degradation and mitotic exit. Piotrowska et al. (Piotrowska, Natorff et al. 2000) reported a homolog of Skp1 protein in *A. nidulans* (sconC).

8. DNA polymerases III subunits tau and gamma identified in *C. immitis* and *F. oxysporum*.

9. Single exon gene in *A. niger* encodes a conserved glutaredoxin domain (DUF836). This domain is known to be present in other *Aspergillus* species (e.g. *Aspergillus clavatus*).

Proteins predicted in F. oxysporum

Among sixteen fungal species in *F. oxysporum* the number of predicted genes predicted by the new algorithm as well as the difference between number of genes in annotation and predictions is the highest (Table 5.1). Therefore, it is not surprising that the umber of the newly identified genes is high for this species.

1. CenPB domain containing protein which binds the centromere and involved in chromosomal stability.
2. Ub-ligase HRD1 associated with Serum-Glucocorticoid-induced Kinase (SGK) degradation.
3. DNA-excision repair protein Rad14.
4. Subunit of ESCRT-II (endosomal sorting complex required for transport II (Hurley 2008).
5. DNA-excision repair protein.

Other functionally characterized genes

1. An interesting protein encoded by a single exon gene is found in *B. cinerea* (Supercontig_1.70, gene_8). It is characterized as a transthyretin precursor; transthyretin is a thyroid hormone receptor also known to act as amyloidogenic protein involved in one of amyloid diseases in humans. Recently, it was found that transthyretin homologs can be found in yeast and even bacteria (Schreiber 2002). A gene for transthyretin precursor is present in annotation of *C. immitis.*
2. A protein product of a predicted four-exon gene is found to be orthologous (no paralogs are detected) to an essential yeast protein Nip7 and a protein in another

Aspergillus species *A. oryzae*; other orthologs of Nip7 are known in eukaryotic species up to mammals. Nip7 is involved in ribosome biogenesis, presumably in rRNA processing regulation.

3. A single exon gene in *A. niger* encodes a conserved glutaredoxin domain (DUF836). This domain is known to be present in other *Aspergillus* species (e.g. *Aspergillus clavatus*).

4. A component of ER-related protein degradation system - DER1 in *B.cinerea*.

5. The actin binding protein profilin in *B.cinerea.*

5.5 Repetitive Sequences in Predictions

For the reasons discussed in Section 4.5 the preliminary masking of input sequences is omitted in this study. Instead, the RepeatMasker program (Smit, R., Green, H., Green P. unpublished work; http://repeatmasker.org) is used to run against the predictions on whole genome level. In general, relatively low volumes of repetitive sequences are identified in all sixteen fungal genomes. The repeat content of the whole genomic sequence varies from 0.2% in *F. verticillioides* to 6.2% in *M. grisea* (Table 5.9). Most of the repetitive sequences identified by RepeatMasker belong to the non-coding states. The exceptions however are *F. oxysporum*, *M. grisea*, *R. oryzae* and *C. cinereus* where the coding sequences contain nearly 60%, 62%, 80% and 82% of all repetitive sequences respectively. Although at first these numbers seem to be high these repeats occupy only 4.6%, 8.9%, 3.7% and 2.0% of all coding sequences for *F. oxysporum*, *M. grisea*, *R. oryzae* and *C. cinereus* respectively (Table 5.9).

Table 5.9 Statistics of the content of repetitive sequences determined by RepeatMasker in protein-coding and non-coding regions (as predicted by GeneMark-ES-2) determined in the sixteen fungal genomes.

species	repeats (nt)				% of all repetitive sequences			repeats found in coding regions as % of total size of predicted coding regions	% of total genome size	genome size (MB)
	in intergenic regions	in coding regions	in introns	total (nt)	in intergenic regions	in coding regions	in introns			
A. nidulans	638,497	186,062	34,227	858,786	74.3	21.7	4.0	1.3	2.8	31
A. niger	141,935	67,487	9,306	218,728	64.9	30.9	4.3	0.4	0.6	34
A. terreus	135,794	18,713	4,159	158,666	85.6	11.8	2.6	0.1	0.5	29
B. cinerea	272,539	100,341	8,753	381,633	71.4	26.3	2.3	0.8	1.5	26
C. immitis	379,493	51,969	54,726	486,188	78.1	10.7	11.3	0.5	1.7	29
C. cinereus	51,643	387,201	24,847	463,691	11.1	83.5	5.4	2.0	1.2	38
C. neoformans	194,860	139,299	27,505	361,664	53.9	38.5	7.6	1.3	1.8	20
F. graminearum	99,662	40,156	2,228	142,046	70.2	28.3	1.6	0.6	0.4	40
F. oxysporum	764,914	1,276,899	88,266	2,130,079	35.9	59.9	4.1	4.6	3.6	60
F. verticillioides	72,123	20,471	4,298	96,892	74.4	21.1	4.4	0.1	0.2	42
M. grisea	783,116	1,518,185	163,183	2,464,484	31.8	61.6	6.6	8.9	6.2	40
N. crassa	682,978	95,603	277,416	1,055,997	64.7	9.1	26.3	0.6	2.7	39
R. oryzae	104,687	660,911	40,280	805,878	13.0	82.0	5.0	3.7	2.0	40
S. pombe	133,175	66,816	14,231	214,222	62.2	31.2	6.6	0.8	1.8	12
S. sclerotiorum	436,924	301,321	25,777	764,022	57.2	39.4	3.4	1.8	2.0	39
S. nodorum	309,373	33,960	31,208	374,541	82.6	9.1	8.3	0.2	1.0	37

5.6 Algorithm Stability with Respect to Random Fluctuations of Gibbs Sampler

GeneMark-ES-2 employs the Gibbs sampling algorithm to align the subset of predicted intron sequences. The alignment is then used to determine the parameters of the BP model as well as the parameters of BP downstream and upstream spacers (Section 5.2.2). The two main steps of the Gibbs sampling algorithm are the predictive update and the sampling step.

The predictive update selects one of the sequences from the input set and places its motif into the background and updates the positional frequency and background models. The point of interest of this section is the second, sampling, step. During this step the new motif position of the sequence, which was selected at the predictive update, is determined by randomly sampling from the weighted motif score distribution. While the highest scoring motif is not guaranteed to be chosen in such an approach, they are more likely to be selected. The problem, however, is that the results with the same input may vary, unless the random number generator uses the same seed, which consequently can alter the final results of GeneMark-ES-2. The question is: "How close are the models and the final genomic parse produced by different runs of GeneMark-ES-2 when applied to the same input sequence?" To address this question ten identical runs of GeneMark-ES-2 are carried in parallel on genomic sequence of *S. pombe*. The splice site and BP motif logos as well as downstream spacer length distributions obtained at the convergence step of each run of GeneMark-ES-2 (Figure A3 of Appendix) show unnoticeable variation. Also, after each run is complete the accuracy of the corresponding final model is evaluated by using *S. pombe*'s test set Type I. Figure 5.20

shows that the intron prediction accuracy levels in terms of Sn and Sp are within the ranges of 0-0.3% and 0-0.2%, respectively.

The variation in the final GeneMark-ES-2 results that are caused by the sampling step is also assessed on the whole genome level. The *S. pombe* genomic parse based on one of the ten models obtained at the convergence of GeneMark-ES-2 run is used as an "annotation".

The performance of the other nine models is evaluated with respect to this "annotation". Figure 5.21 shows that the difference in intron detection on whole genome level reaches 1.3% (Sn) and 1.0% (Sp) between GeneMark-ES-2 run #1 and run #6. On average the difference in the termination site detection with respect to "annotation" is only 0.2% (Sn) and 0.3% (Sp). Furthermore, the total number of predicted genes ranges from 4,822 to 4,834 and the total number of exons varies from 9,770 to 9,804. These results indicate that the main differences between the runs are due to shifts in acceptor and/or donor prediction.

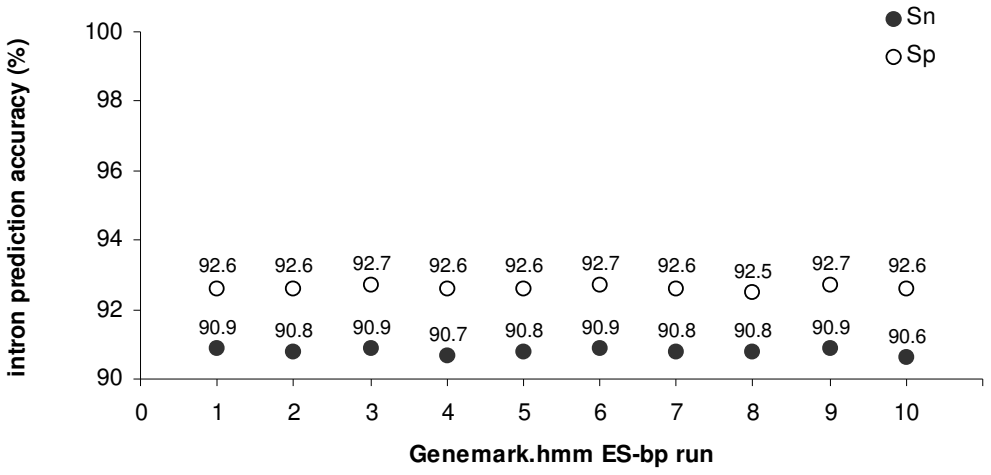


Figure 5.20 Intron prediction accuracy values for ten models produced by GeneMark-ES-2 run on the genomic sequence of *S. pombe*.

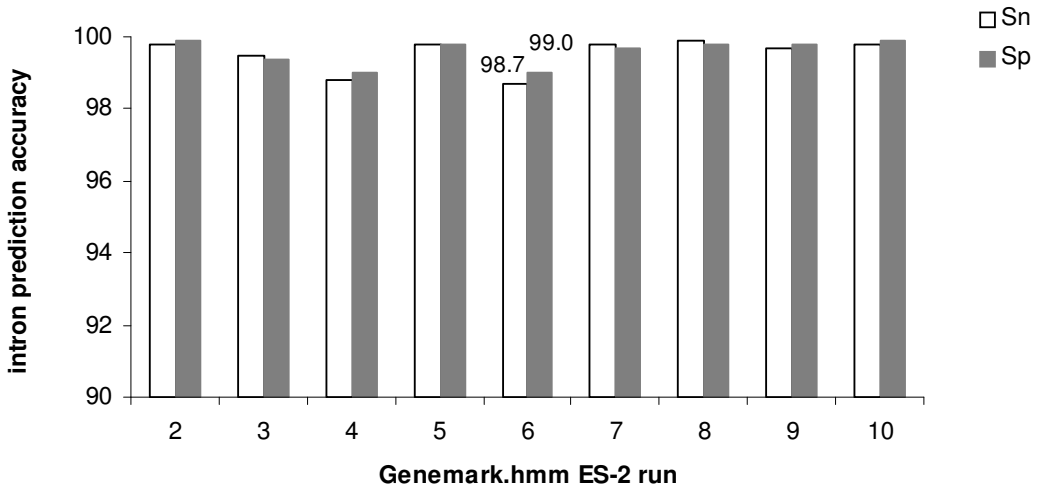


Figure 5.21 Intron prediction accuracy values for nine models produced by GeneMark-ES-2 run on the genomic sequence of *S. pombe*. The genomic parse of the first GeneMark-ES-2 run is used as annotation in accuracy determination. The uppermost deviation from the annotation is observed for run #6.

5.7 Conclusions

An automatic *ab initio* gene prediction algorithm GeneMark-ES-2 for fungi is introduced in this chapter. The algorithm uses new intron submodel to better describe gene organization of fungal genomes. It employs an iterative training strategy in which the HMM parameters and its architecture are automatically adjusted to reflect genome-specific properties, such as splicing mechanism or exon-intron organization. The noteworthy advantage of GeneMark-ES-2 is its ability to generate accurate gene models without a prior knowledge characterizing species under study (e.g. training set).

The algorithm was applied to the genomes of sixteen publicly available fungal species. The results indicate that GeneMark-ES-2 predicts genes with higher accuracy than the original self-training algorithm and other algorithms which utilize supervised

training; the algorithm predicts splice sites in all sixteen fungal genomes with higher than 85% accuracy (Sn and Sp). Furthermore, the proposed method demonstrates improvement over current annotations as tested on genomes which exhibit symptoms of over-annotation or under-annotation.

An *ab initio* gene finder with unsupervised parameter estimation provides flexibility and practical applicability in the annotation process. In early stages of genome sequencing project when a validated set of genes is not available the self-training algorithm, currently, is the only *ab initio* tool that can successfully be used.

GeneMark-ES-2 is part of the annotation pipeline in several sequencing centers including Broad Institute and JGI; the software package is also publicly available for non-commercial use (Chapter 1). For the genomes which contain a small number of introns, such as low eukaryotes, the amount of data necessary for accurate description of exon-intron structure becomes insufficient. This problem is addressed in the next chapter which discusses the gene prediction in low eukaryotes.

CHAPTER 6

GENE FINDING IN GENOMES WITH SMALL NUMBER OF INTRONS

6.1 Introduction

The high degree of diversity in more than 1,000 eukaryotic genomes that are either completely sequenced or under sequencing is indisputable. Even within the same *order* the genomes exhibit significant evolutionary divergence. Genomes analyzed in Chapter 5, for example, are populated with genes containing on average from one to five introns per gene (see Table 5.1) The development of one "universal gene prediction algorithm" applicable to any type of eukaryotic genome is currently a challenging task. The new unsupervised training method, GeneMark-ES, is developed for eukaryotic species which contain introns with a long poly-Y tail upstream of acceptor site as well as a weak BP signal (see Chapter 4). The new intron submodel utilized in GeneMark-ES-2 is introduced to better characterize genomes with genes which possess introns with no poly-Y tail and with strong BP motif.

Yeast-like genomes represent a category of eukaryotic genomes with a small fraction of intron containing genes. In such genomes detection of single exon genes can be carried out with high accuracy by prokaryotic and eukaryotic gene finders. For instance, application of GeneMarkS to Chromosome IV of the *S. cerevisiae* genome shows 96.2% Sn and 91.6% Sp in translation termination site prediction accuracy and 84.2% in exact gene prediction accuracy as tested on the set of annotated genes. An

accurate prediction of multiple exon genes in such genomes is a challenging problem for algorithms using unsupervised training. The main difficulty is to pinpoint a sufficient number of spliced genes to derive model parameters reflecting exon-intron structure.

Frequently, the problem of exon-intron structure identification in Yeast-like genomes is addressed by the use of the gene models from well-studied and closely related species (e.g. *S. cerevisiae* model is applied to the genomic sequence of *Saccharomyces paradoxus*). Such an approach is likely to fail as the evolutionary distance between species increases. The unsupervised training algorithm, GeneMark-ES-2 performs well when applied to fusion yeast, *S. pombe*. The *S. pombe* genome contains on average one intron per gene (Table 5.1), thus a subset of about 5000 introns can be used for training. Nevertheless, as the number of spliced genes in a genome decreases the efficiency of GeneMark-ES-2 is expected to decline.

This chapter describes a semi-supervised gene prediction algorithm for eukaryotic genomes with a small number of introns. The algorithm, GeneMark-LE, is applied to the completely sequenced genomes of *Hemiascomycetes* (Figure 6.1). The results show that the algorithm produces satisfactory accuracy of exon-intron boundary prediction. The models derived by the algorithm on average perform better than the gene models transferred from a closely related and well characterized species. GeneMark-LE can be applied to genomes that have already been annotated and are in the process of re-annotation efforts as well as to the genomes that are being sequenced.

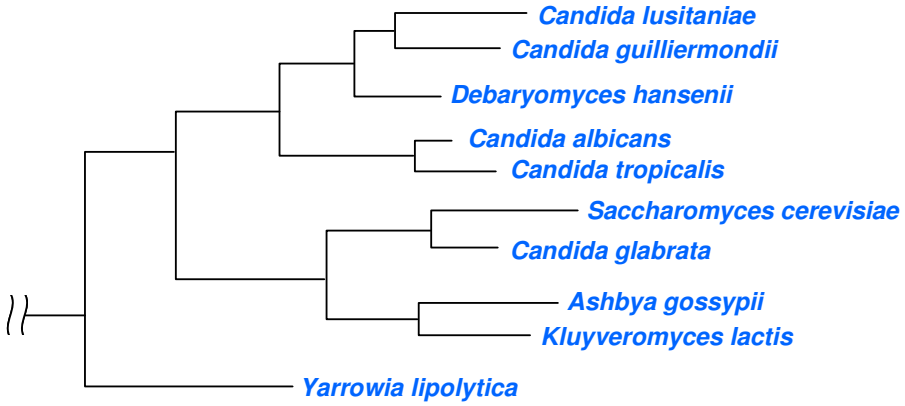


Figure 6.1 The phylogenetic relationships within fungal class of *Hemiascomycetes*. The tree is based on the sequence data obtained from http://fungal.genome.duke.edu.

6.2 Motivation for semi-supervised gene model

Testing of GeneMark-ES-2 demonstrated that the introduction of new intron submodel led to a significant improvement in algorithm performance (Chapter 5) if a fungal genome contains on average one to five introns per gene. The question is: "What is the minimum number of introns per gene in a given genomic sequence sufficient for successful application of GeneMark-ES-2?" To address this question the algorithm could be applied to several genomes which fall between *S. cerevisiae* and *S. pombe* in terms of average number of introns per gene. In addition, an adequate numbers of EST sequences or experimentally verified multiple exon genes are required to obtain a test set to assess the GeneMark-ES-2 accuracy. Unfortunately, such genomes presently are not available. Still, it is possible to fill this gap by generating artificial genomes. Two different approaches can be implemented to perform this task. The first approach includes randomly masking of the predefined fraction of introns in annotation of a well-known

genome (e.g. *S. pombe*) and utilizing the masked genomic sequence as an input to GeneMark-ES-2. This experiment can then be repeated several times by changing the fraction of masked introns. One of the drawbacks of this approach is that the annotation does not necessarily have coordinates of the whole set of true introns, thus the number of introns after masking is not known. Furthermore, the masking procedure may force the algorithm to go through less likely sequence of HMM states e.g. splitting a gene by labeling an intron as intergenic region.

The second approach dynamically reduces the training data corresponding to the set of states presented in Figure 5.3 by the predefined fraction. The negative aspect of this method is the implementation of a dynamic scheme. In contrast to the first approach, where in each iteration the coordinates of the masked data do not change, the second approach supplies data with different content and size. For example, assume that at a given iteration *N*, a total of *X* introns are predicted. The set *X* most likely does not exactly match with the set of introns predicted at iteration *N-1*. Predefined fraction of introns is randomly selected from the set *X* for training. As a result this subset of introns is quantitatively and compositionally different from that of obtained at iteration *N-1*. Another problem associated with both types of simulations is that the input sequence belongs to one and the same genome and does not reflect the variability of genome organization.

Each of the approaches has its own advantages and disadvantages but given the importance of knowing the number of spliced genes in the input sequence or in the training, the second approach is more suitable for the given task. The dependence of average prediction accuracy of internal exons, characterized by (Sn+Sp)/2, on the number

of introns in the training set is shown in Figure 6.2. GeneMark-ES-2 shows reasonable accuracy (87.5%) after training on as little as 700 introns. This number is set as the lower bound above which GeneMark-ES-2 can be applied. Note that the expected number of introns in a new genomic sequence may be known *a priori* or can be estimated from the run of the self-training algorithm.

Notably, the minimum number of 700 introns is an estimate; for genomes which differ in their exon-intron structure from that of *S. pombe* it may vary. Nevertheless, the Yeast-like genomes typically contain less than 700 introns. For these genomes a new semi-supervised training approach is developed which utilizes the data from the native genome as well as from well-studied closely related species.

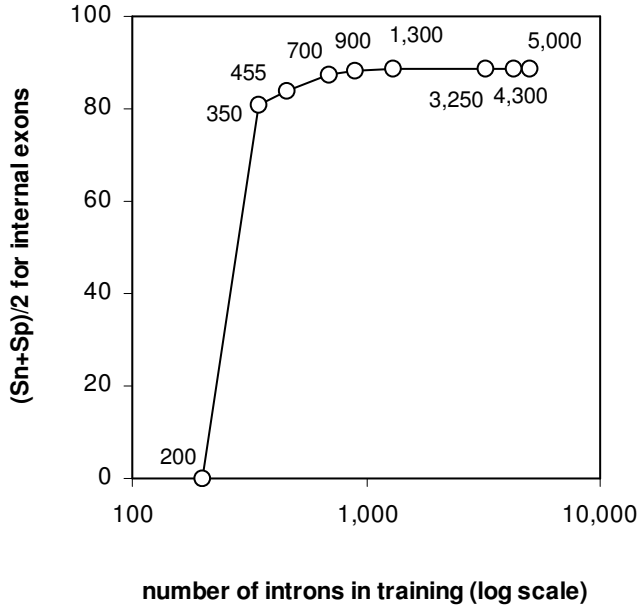


Figure 6.2 The average prediction accuracy of internal exons by GeneMark-ES-2 as a function of number of introns in the training set. The number of introns in the final training set is estimated from the final step of the algorithm.

6.3 Methods

This section describes the semi-supervised gene finding algorithm for fungal species with a small number of introns. The architecture of the underlying HMM is adopted from GeneMark-ES-2. The step-wise diagram of the algorithm is shown in Figure 6.3. It employs (i) iterative unsupervised training (gray arrows) to derive species specific model (U-Model) parameters for coding and non-coding states; and (ii) supervised training to provide initial model (S-Model) parameters required to describe the states associated with intron splicing (e.g. donor and acceptor site models, BP model, spacer duration).

6.3.1. *U-Model Parameter Estimation*

Genomic sequences of ten *Hemiascomycetes* (Table 3.1) species exhibit significant variation in their average G+C content (Table 3.1). The difference of 12% in average G+C content is observed within species of the *Candida* genus (*C. tropicalis* (33%), *C. lusitaniae* (45%)). Similar divergence in average G+C content of genomic sequences could be expected to be present in other closely related species as the genomic sequence for such organisms becomes available. Note that these differences reflected in G+C content of genomic sequences indicate divergence in their codon usage patterns. Hence, the species specific Markov models describing the coding and non-coding regions utilized by U-Model states are essential for accurate gene prediction. In the first run the algorithm is using the heuristic model described in Section 4.2.1. The set of predicted genes is reduced by removing the protein coding regions with length less than 300 nt.

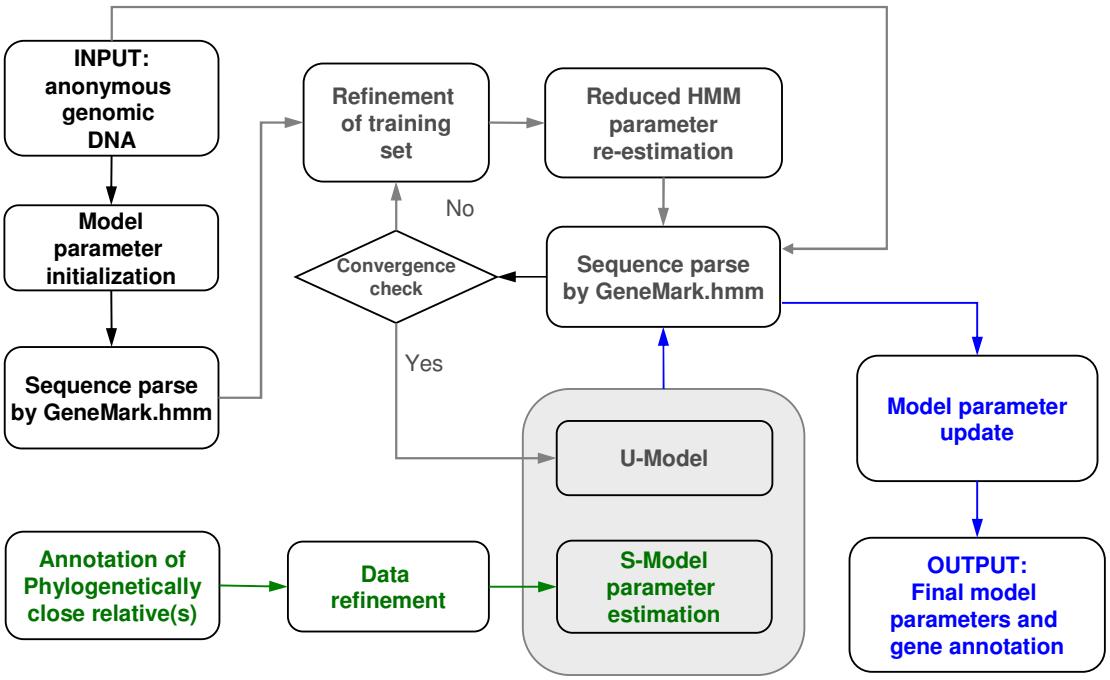


Figure 6.3 Block diagram of GeneMark-LE, a semi-supervised gene finding algorithm for genomes with a small number of introns. The predictions of the heuristic model are used to derive model parameters for U-Model (U-unsupervised) as described. S-Model (S-supervised) parameter estimation is done by using the annotation of closely related species.

The refined training set is used for parameter re-estimation of the models of coding and non-coding regions as well as the models of translation initiation and termination sites. In Yeast-like genomes the spliced genes as well as the splice sites are present in limited numbers. As it was shown in Section 6.2 the parameter update for intron model implemented during iterations did not result in accurate prediction of introns. Therefore, we use reduced HMM architecture in which only models of coding and non-coding regions as well as models of translation start and stop sites are updated.

6.3.2. *S-Model Parameter Estimation*

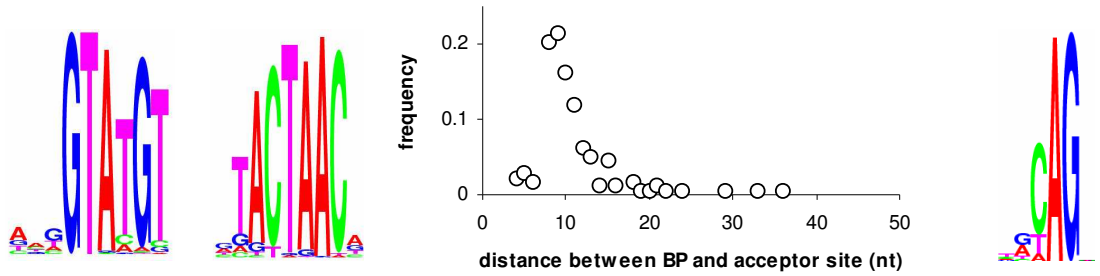
While the G+C content and subsequently the codon usage vary in closely related species the signals involved in intron splicing mechanism are rather conserved (Figure 6.4). With the exception of *Y. lipolytica*, which to be discussed later in this section, the donor site signal is well conserved in all species. The strict consensus "GT" of the donor site is followed by a sequence with consensus "ATGT". Similarly the BP signal is conserved among these species. An acceptor site sequence which does not carry significant information also exhibits a signal common to fungal species. Hence, the S-Model is designed to capture the shared splicing properties of the Yeast-like genomes.

It should be noted that the intron annotation data, which is obtained from Ares lab (www.cse.ucsc.edu/research/compbio/yeast_introns.html) and used to produce results for *S. cerevisiae* shown in Figure 6.4 is supported by experiments. For other genomes the NCBI annotation is used for this purpose. The introns are extracted from the set of multiple exon genes to generate logos for donor and acceptor sites. Gibbs sampling algorithm is then used to identify the BP motif and its downstream spacer length distribution from the set of these introns. Therefore, the marginal differences in the splice site frequencies as well as the relatively lower frequencies observed in BP motif compared to that of *S. cerevisiae* are not surprising. In contrast to the models of sites the shapes of downstream spacer length distribution demonstrate noticeable differences. Five species exhibit low (*K. lactis* and *S. cerevisiae*) moderate (*C. albicans*) and high (*A. gossipii* and *D. hansenii*) downstream spacer length localization. Interestingly, however, the *A. gossipii* and *K. lactis* are the closest relatives among ten species (Figure 6.1). Section 6.3.3 addresses the issue of the genome specific state durations.

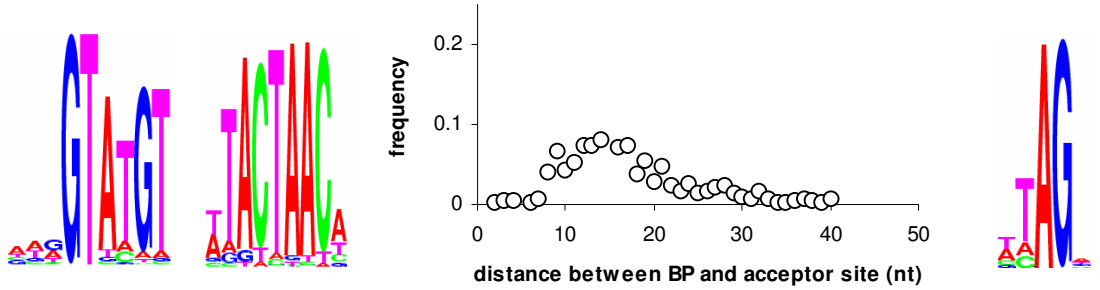
The data used for S-Model parameter estimation is initially pre-processed to remove (i) non-canonical signal sizes, (ii) exon/introns with lengths shorter than 30 nt, and (iii) exons with in-frame stop codons. Experimentally verified introns of *S. cerevisiae* (Section 3.3) are used to derive the (i) parameters of the models associated with signal site states i.e. donor, acceptor, and BP sites; and (ii) state durations for introns and BP downstream spacer. The exon length distribution as well as state transition probabilities such as *P(bp)* and *P(ie)* (Figure 5.3) are derived from *S. cerevisiae* annotation. The annotated *S. cerevisiae* genes contain only nine internal exons. Estimation of state duration from only 9 data points is not practical. Therefore, the length distribution for this state is modeled from whole set of exon lengths.

Significant difference in splice site organization is observed for *Y. lipolytica*. This species has been labeled as "non-conventional" yeast species based on its genetics, physiology and phylogeny (Barth and Gaillardin 1997). The consensus "GT" in the donor site is followed by a sequence with strongly conserved "GAGT". Moreover, the position of BP site motif is highly localized with 80% of consensus sites located only 2 nt upstream of acceptor site (Figure 6.4). These findings emphasize the distance of *Y. lipolytica* from other *Hemiascomycetes* indicating a difference in evolutionary processes that shaped splicing mechanisms of these species. Therefore, to provide training and test sets for *Y. lipolytica* the annotated 672 introns containing genes were randomly split into test and training sets in 1:2 ratio, respectively.

A. gossypii



C. albicans



D. hansenii

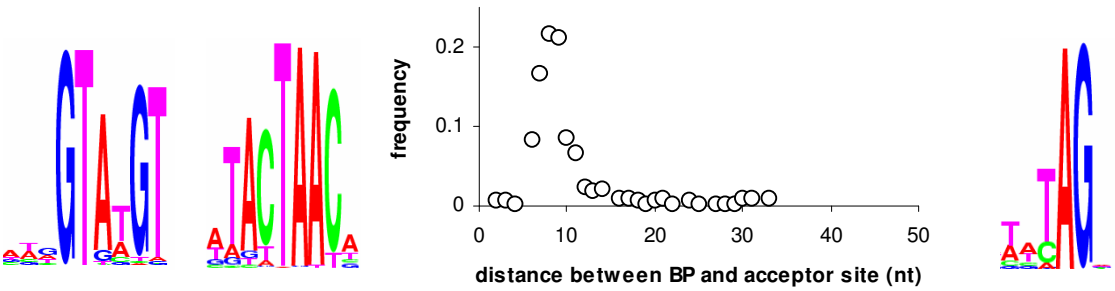
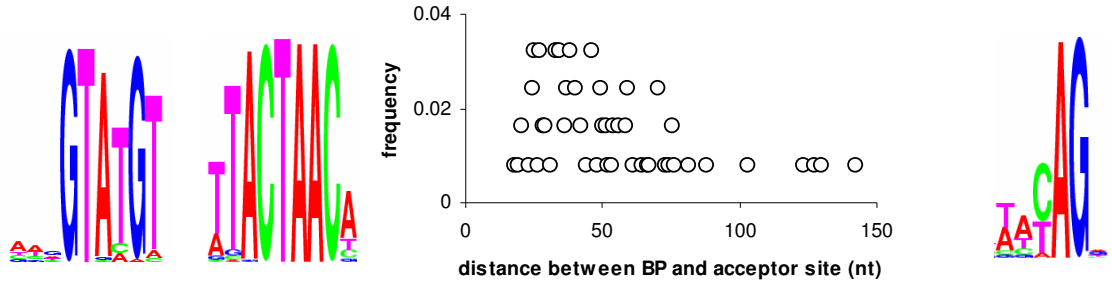
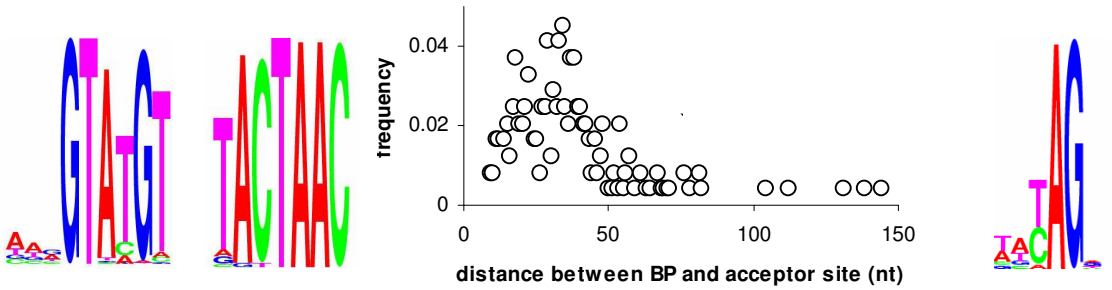


Figure 6.4 Logos for donor, acceptor and BP sites as well as the downstream spacer length distributions. An experimentally validated set of introns is used for *S. cerevisiae*. For other species the set of annotated spliced genes is used. High degree of divergence is observed for the BP downstream spacer length distribution.

K. lactis



S. cerevisiae



Y. lipolytica

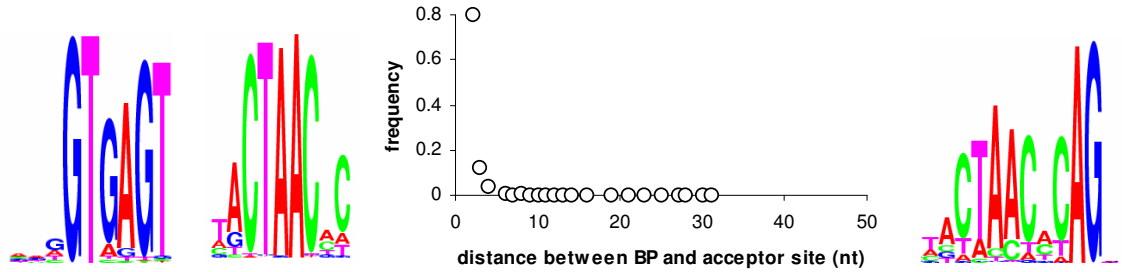


Figure 6.4 (continued)

6.3.3. *Semi-Supervised Model*

A hybrid model is generated by combining corresponding model parameters from U-Model and S-Model (Figure 6.3 blue arrows). U-Model contribution to the hybrid model parameterization includes (i) the 5th order inhomogeneous and 5th order inhomogeneous Markov models describing the coding and non-coding states respectively; (ii) the state durations for single exon genes; and (iii) the positional

frequency models of translation initiation and termination sites. S-Model provides models for (i) donor, acceptor and BP sites; (ii) the length distributions for exons (excludes single exons); and (iii) length distributions of introns and BP upstream and downstream spacers. The hybrid model is then used to parse the input sequence (Figure 6.3) with GeneMark.hmm E-3.0. Finally, the semi-supervised model is derived from this parse thus allowing for tuning the model to better characterize the sequence patterns existing in the genome in question.

6.4 Results and Discussion

The sets of selected annotated genes (Section 3.4.3) are used to assess the algorithm accuracy in terms of Sn and Sp (Table 6.1). Overall, the results show high accuracy of gene prediction. The average intron prediction accuracy for all five genomes is above 86%, reaching 90.3% for *C. albicans*. The termination site prediction accuracy, observed highest for *A. gossypii* (98%), shows 95% and higher for all species indicating that in general the possible mislabeling of genomic sequences does not affect the reading frame of the gene.

Further rounds of iterations, with the semi-supervised model used as a starting point, do not result in noticeable changes in accuracy results. In the case of *A. gossypii*, for which one of the highest gains in accuracy was observed, the next iteration shows only 1% difference at nucleotide level and 2% in stop codon prediction compared to the semi-supervised model. The difference in exon prediction accuracy observed on the test set is only 0.3%.

Table 6.1 Sensitivity and specificity (Sn/Sp) values and their average for several categories of gene structure accuracy of gene prediction of GeneMark-LE.

		A. gossypii		*C. albicans*		*D. hansenii*		*K. lactis*		*Y. lipolytica*	
Intron	Sn	87.3	88.8	87.6	90.3	88.7	90.0	86.1	89.3	83.1	86.1
	Sp	90.3		93.0		91.3		92.5		89.1	
Donor	Sn	94.7	96.3	92.7	95.6	91.5	92.9	90.3	93.7	86.3	88.9
	Sp	97.9		98.4		94.2		97.0		91.4	
Acceptor	Sn	89.3	90.6	88.0	90.7	90.8	92.2	87.5	90.8	86.7	89.7
	Sp	91.8		93.4		93.5		94.0		92.6	
Exon	Sn	88.0	88.6	88.8	89.4	87.4	88.1	84.7	85.9	79.7	80.8
	Sp	89.2		89.9		88.7		87.1		81.9	
Initiation site	Sn	87.2	87.2	95.4	93.7	90.4	90.4	86.1	85.5	76.3	75.8
	Sp	87.2		92.0		90.4		84.9		75.3	
Termination site	Sn	99.3	99.0	98.3	96.6	95.6	95.6	97.2	96.6	96.0	95.8
	Sp	98.7		94.8		95.6		95.9		95.6	
Nucleotide	Sn	98.6	99.1	98.6	99.0	98.7	98.6	97.0	97.4	97.0	97.3
	Sp	99.6		99.4		98.5		97.7		97.6	

While the performance of GeneMark-LE, as determined on the test sets, is satisfactory a question can be posted: "What is the advantage of employing such a training approach as opposed to using gene models from closely related species e.g. *S. cerevisiae*?" Another question is: "What is the importance of the final step of the algorithm which determines the semi-supervised model (blocks connected by blue arrows in Figure 6.3)?" To address these questions final model performance is compared to the performance of the two following models.

The first model is the native *S. cerevisiae* model, which includes the S-Model, and the second is the hybrid model based on the combined U-Model and S-Model parameters. The parameter estimation process in each case is carried out by the same procedure. Figure 6.5 shows the absolute difference between the accuracy values of the semi-supervised model and the Yeast derived model (left panel of the Figure 6.5) as well as the

accuracy difference between the semi-supervised and hybrid models (right of the Figure 6.5). The semi-supervised model outperforms both models for all species. Significant gain in accuracy is observed for *D. hansenii* and *A. gosypii*.

The difference in Sn values for these species is more than 33% as compared to both the Yeast and the hybrid models. The Sp gain is significant for *D. hansenii* in both comparisons (above 23%). For *A. gossypii* the improvement in Sp values is more than 8%. The difference in accuracy of intron prediction is not as dramatic in the cases of *C. albicans* and *K. lactis*. For these species the semi-supervised model shows higher Sn values by 2.8% and 4.6% respectively. Semi-supervised model shows slightly better specificity results for *K. lactis* and *C. albicans* (Yeast model). A marginal decrease (-0.9%) in intron prediction specificity is observed for *C. albicans* as the semi-supervised model is compared to the hybrid model. The large variation of results observed in Figure 6.5 is not surprising. As stated above, *A. gossypii* and *D. hansenii* exhibit significant divergence in their BP downstream spacer length distribution from the one observed for *S. cerevisiae*.

The location of BP site is less localized in *C. albicans*. *K. lactis*' downstream spacer length distribution shows a shape similar to that of *S. cerevisiae*. The results underline the importance of use of native (semi-supervised) model even for closely related species.

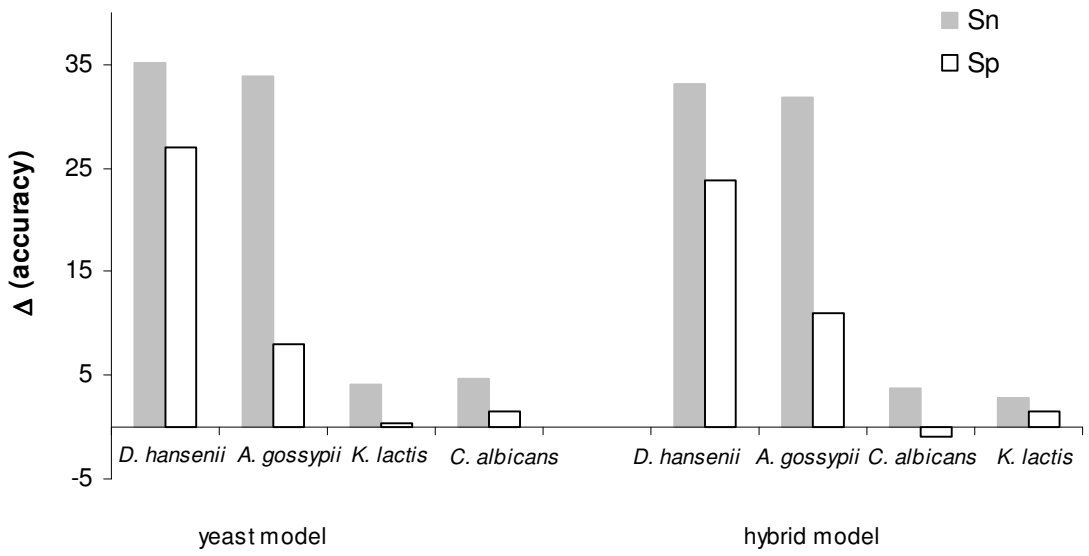


Figure 6.5 The difference in prediction accuracy of introns between the semi-supervised models and yeast model (left), and between semi-supervised and hybrid models.

Yeast-like species show distinctive differences between initial and terminal exon lengths (Figure 6.6). Generally, the initial exons are skewed towards the shorter lengths. Terminal exons have less localized distribution exhibiting mean exon lengths longer than that of initial exons.

6.5 Conclusions

GeneMark-LE is developed for eukaryotic genomes with a small number of introns. The results indicate that the semi-supervised model outperforms the models which are based on the data obtained from closely related species (*S. cerevisiae*) as applied to *A. gossypii*, *D. hansenii*, *C. albicans* and *K. lactis*. For species closely related to *S. cerevisiae* the initial model parameters describing states involved in intron splicing (S-Model) are derived from experimentally verified data of *S. cerevisiae*. For species

exhibiting a closer phylogenetic relationship to *Y. lipolytica*, which demonstrates significant difference in its splicing properties as compared to *S. cerevisiae,* the S-Model parameters are derived from *Y. lipolytica*.

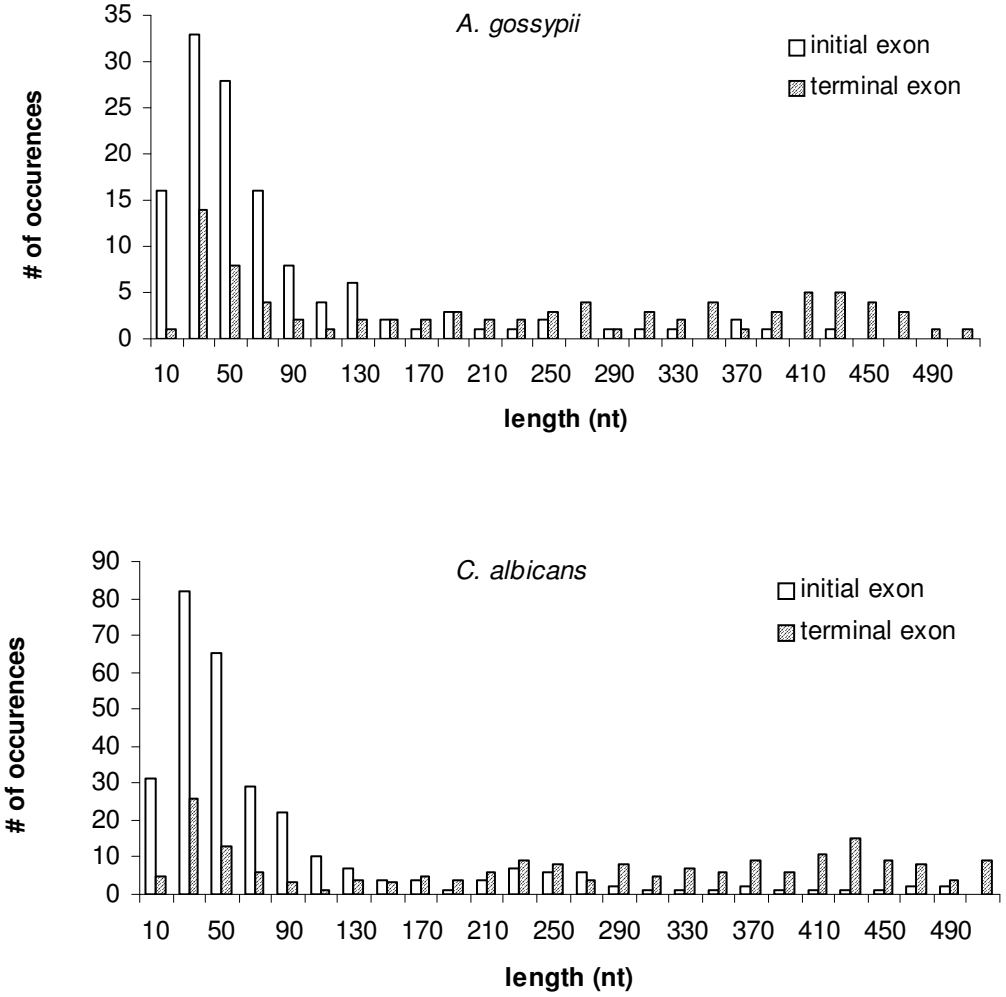
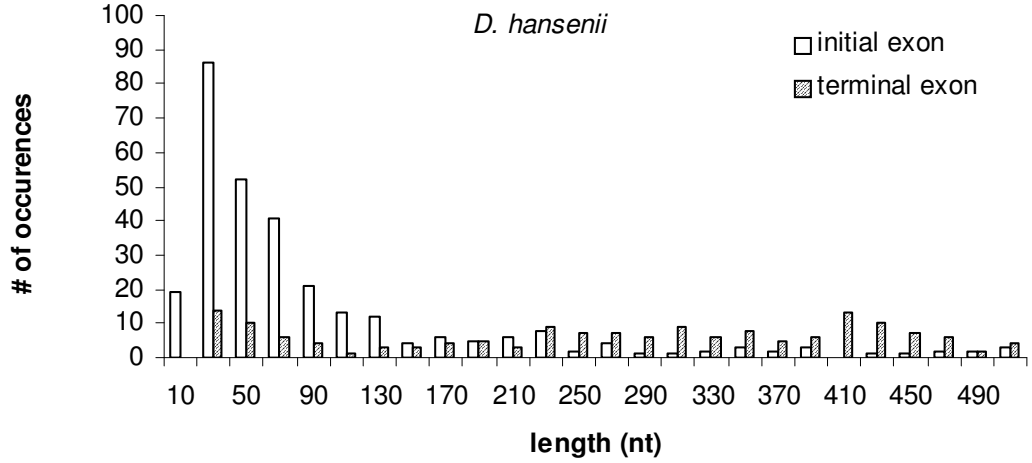
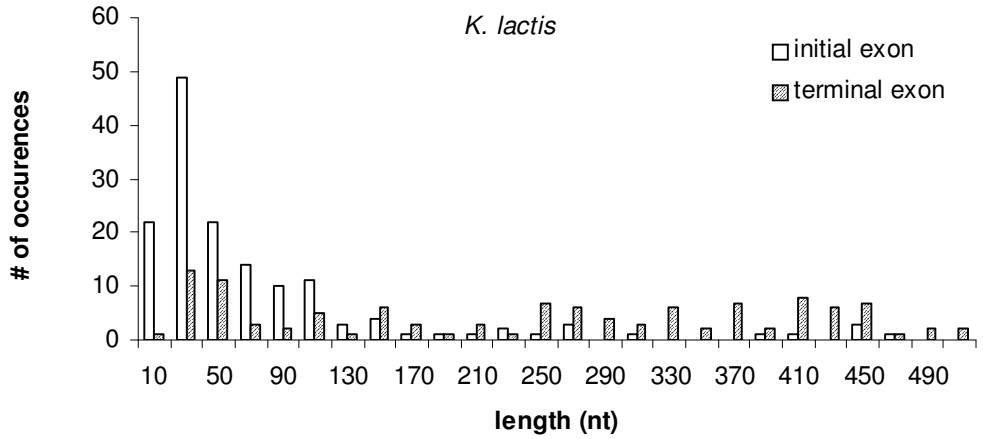


Figure 6.6 Histograms of initial and terminal exon lengths obtained from the predictions of the final GeneMark-LE model.





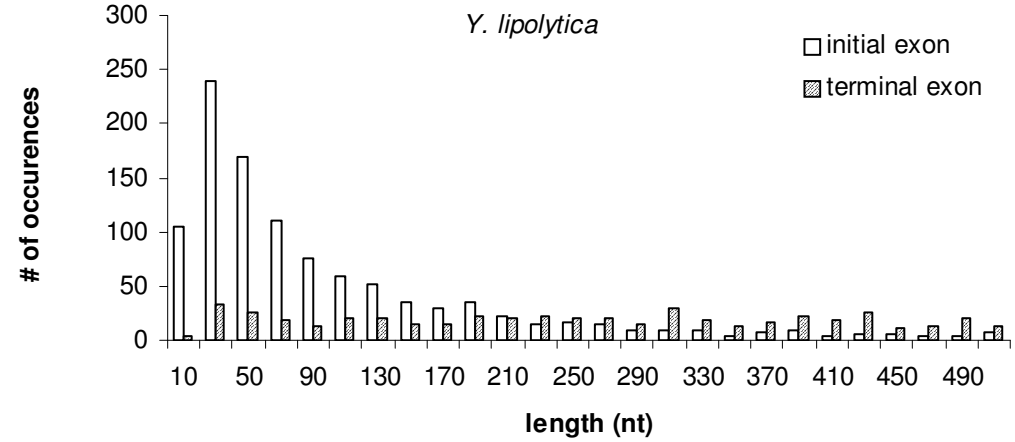


Figure 6.6 (continued)

CHAPTER 7

CONCLUDING REMARKS

7.1 Current Research

Three types of training algorithms for eukaryotic gene finding are described in this thesis. Each of these algorithms is designed for better performance within a distinct class of eukaryotic species. GeneMark-ES, an unsupervised approach described in Chapter 4 is developed to be applied to eukaryotic species with intron sequences possessing strong donor and acceptor sites, long poly-Y tail upstream of acceptor site, and weak BP site. Introduction of new intron submodel in GeneMark-ES-2, described in Chapter 5, allows better characterization of gene structure of the fungal genomes. Finally, the GeneMark-LE is designed for gene prediction in Yeast-like genomes that contain a small number of spliced genes.

The important feature of this research is making the generation of reliable gene models for gene prediction independent from the process of data validation necessary for training of the traditional *ab initio* gene finding methods. The practical importance of the algorithms is unquestionable given the abundance of genomic sequences produced at ever increasing rate.

7.2 Major Challenges in Unsupervised Training Procedure

The species described in this thesis vary in their phylogenetic and genomic characteristics e.g. G+C content, splicing properties and size. The eukaryotic genomes in

general can also be classified as species containing low, medium, and high volumes of non-coding sequence (the Yeast, the worm, and the human genomes). As the size of non-coding sequence grows the noise level in training procedure increases as well; it imposes not only computational strains on the algorithm, but also negatively affects the discrimination power of the statistical models. Transposable elements, simple repeats and other repetitive elements are present in abundance in plants and vertebrate genomes (Lander Linton et al. 2001; Feschotte, Jiang et al. 2002; Consortium 2005).

Gene duplication events, can be considered as repetitive families from the gene finding standpoint and they may significantly bias the model parameters of protein coding regions.

In addition to repetitive elements in higher eukaryotes the compositional characteristics of the genomic sequence become more complex as well. The nhomogeneous distribution of DNA composition known as isochore theory (Cuny, Soriano et al. 1981; Bernardi 2000) raised questions about the functional significance and the underlying evolutionary processes behind these mosaic patterns. With respect to gene finding this means that the genes which belong to different isochores also vary in their codon usage. Consequently, derivation of a single gene model to predict genes in all G+C clusters is not a promising approach.

Current *ab initio* gene finders that employ supervised training experience difficulties when applied to higher eukaryotes (Guigo, Flicek et al. 2006). Hence, expectation of performance similar to that of described in Chapters 4 and 5 in this case is not justified.

7.3 Future Directions

The future research could include development of modules addressing set of problems stated in the previous section.

Masking of the input sequence is one of the essential steps that should be included in the input sequence pre-processing procedure (see Section 4.2.7). To avoid under-masking or over-masking (see Section 4.5) a more comprehensive approach sequence should be developed. In the training set refining process the predicted genes which are adjacent to the repetitive sequences should be either considered with caution before they are included in the training set, or removed from consideration.

In order to account for the codon usage differences of the genes stipulated by the isochors the modeling of gene elements with different G+C content is one of the necessary steps in describing the gene structure in higher eukaryotes. In addition to modeling step the prediction step should be modified as well. Evaluating the gene prediction accuracy for supervised model parameterization Sparks and Dorman demonstrated that the input sequence segmentation into various G+C regions in parallel with test set segmentation leads to improvements in gene prediction accuracy in rice (Sparks and Dorman 2007). The input sequence segmentation into corresponding G+C clusters follows by the iterative application of the gene finder with a gene model describing the particular G+C range. The difficulties occur in boundaries where neighboring G+C segments join. The calculation of the *a posteriori* probability for the region that contains transition from one G+C cluster to another is relatively straightforward procedure which demands computational adjustments e.g. automatically switching the gene models in the boundaries. Yet the determination of the values of state

durations is a difficult task since the simple plug-in of values violates the normalization conditions. It is possible to circumvent this problem by simply deriving the state durations from the whole set of observations instead of the corresponding G+C cluster. Otherwise, the values of state durations should be dynamically adjusted during the Viterbi parse of the algorithm.

Algorithm modifications and integration of the additional modules, however, may lead to the term "ab initio" to be omitted. Moreover, the label self-training in its pure form can be argued since *a priori* information such as coordinates of repetitive elements and/or GC isochors is supplied to the algorithm.

Integration of external information from databases of ESTs, cDNAs and proteins, as well as syntheny regions between genome in question and a closely related species can improve the self-training algorithm performance. The extrinsic information can be utilized (i) in the prediction step of the algorithm where the *a priori* known coordinates of exon-introns boundaries are used as hints and (ii) in the step of training set derivation where the known (labeled) sequences can be used to update the models of protein coding and non-coding regions, splice sites, and length distributions.

Extrinsic information is the most useful when it is verified by experiment. The difficulties arise when the supplied information contains high level of noise reflected in contaminated EST databases, false positive hits, or introns in UTRs (see Chapter 2 for more details). The complete gene transcripts obtained from EST-to-DNA alignments should be further for a presence of uninterrupted (no-frame shifts) protein coding regions. Similar approach is applied in deriving the test set for assessing the algorithm accuracy (Section 3.4.1).

APPENDIX

SUPPLEMENTARY FIGURES AND TABLES

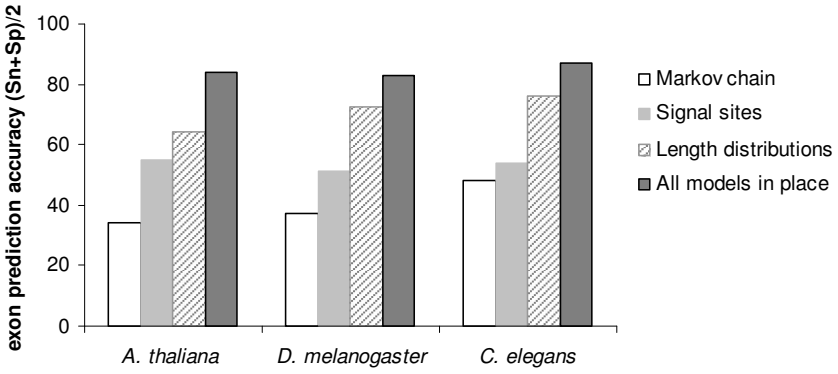


Figure A1. Exon prediction accuracy with a particular submodel used in its minimal form. The results are obtained from supervised training model

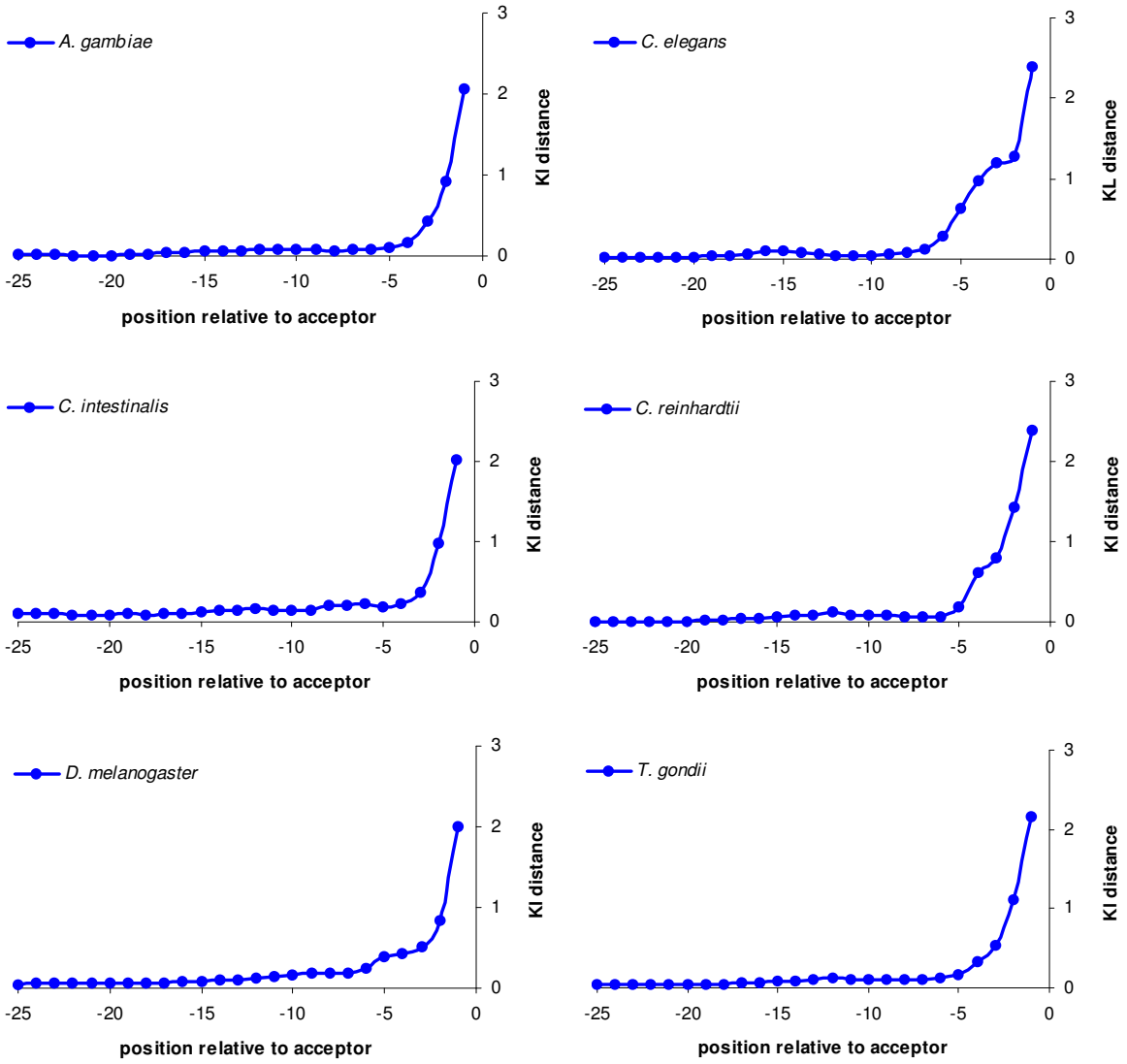


Figure A2. Kullback-Liebler distance for upstream of acceptor site (excluding the acceptor site) for six eukaryotic species. The graphs were generated from last iteration of self-training algorithm applied to a particular species. Results for *A. thaliana* are shown in Figure 4.2 (Section 4.4).

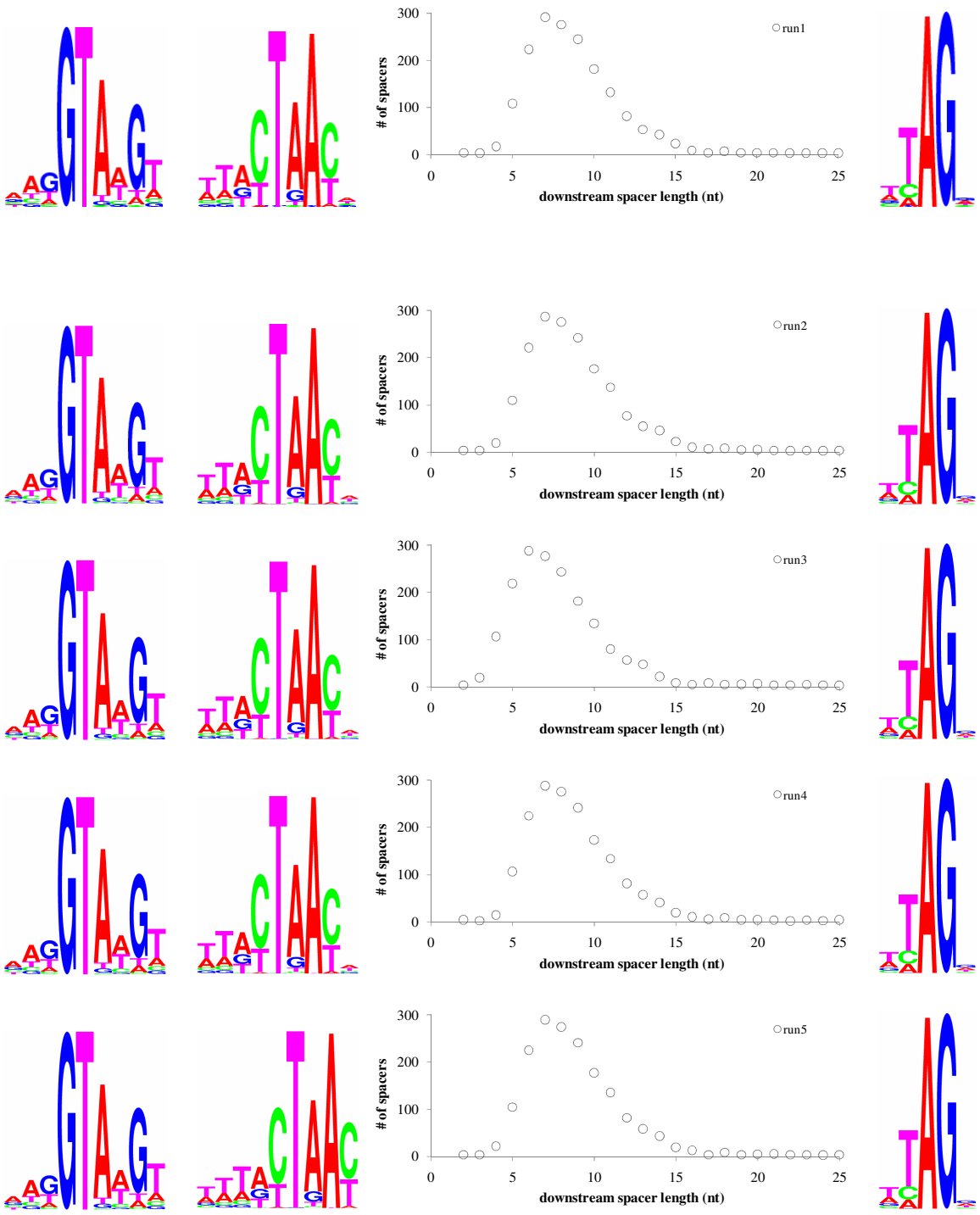


Figure A3. Characteristics of elements of splicing mechanisms for five (out of total 10) runs of GeneMark-ES-2. Similar results are obtained for other runs.

Table A1. List of genes predicted by GeneMark-ES which have a hit to a domain in CDD and are missed in annotation.

#	# of exons	length (aa)	Domain Name	Description of the function
a) *A. thaliana*				
1	1	115	INT_IntI	IntI (E2) integrases, site-specific tyrosine recombinases, DNA breaking-rejoining enzymes, N- and C-terminal domains. This CD includes integrases which are components of multiresistant integrons and mediate recombination between a proximal attI site and a secondary target called the attC.
2	1	219	CAT	Chloramphenicol acetyltransferase.
3	1	385	Soj	Involved in chromosome partitioning and cell division.
4	14	616	AsnB	Asparagine synthase (glutamine-hydrolyzing.) Amino acid transport/ metabolism.
b) *A. gambiae*				
1	2	73	REX1	DNA repair REX1 is required for DNA repair in yeast, and has homologues in other eukaryotes.
2	2	78	COX6C	C-oxidase subunit Vic. A 13 sub-unit complex, EC:1931 is the terminal oxidase in the mitochondrial electron transport chain. This family is composed of cytochrome c oxidase subunit Vic.
3	2	83	Complex1_LYR	1 protein (LYR family). This family of short proteins includes proteins from the NADH-ubiquinone oxidoreductase complex I. The family also includes the B14 subunit from Cow, and the B22 subunit from human.
5	2	89	UPF0239	Protein family (UPF0239)
6	2	108	CHCH	Conserved motif in the LOC118487. Contains Cox19 codes for an 11-kDa protein (Cox19p) required for expression of cytochrome oxidase.

Table A1 (continued)

#	# of exons	length (aa)	Domain Name	Description of the function
7	3	116	Ribosomal_L44	Protein L44.
8	1	222	RimI	General function prediction only.
9	1	227	LITAF	Membrane-associated motif in LPS-induced tumor necrosis factor alpha
10	2	229	Euk_Ferritin	Ferritin (Euk_Ferritin) domain. Ferritins are the primary iron storage proteins of most living organisms and members of a broad superfamily of ferritin-like diiron-carboxylate proteins.
11	1	234	COG4934	Protease Posttranslational modification, protein turnover, chaperones.
12	4	324	Transposase_1	This family includes the mariner transposase.
13	3	330	MRJP	Royal jelly protein; also the sequence-related yellow protein of drosophila which controls pigmentation of the adult cuticle and larval mouth parts.
14	3	364	TAP42	Involved in regulation of TOR signaling pathway.
15	1	390	PgsA	Synthase Lipid metabolism.
16	1	470	COG5062	Membrane protein Function unknown
17	3	496	Amino_oxidase	Contains amine oxidoreductase. This family consists of various amine oxidases, including maze polyamine oxidase (PAO) and various flavin containing monoamine oxidases (MAO).
18	2	583	EGL-9	Proline hydroxylase Posttranslational modification, protein turnover, chaperones .
19	1	733	Neurochondrin	This family contains several eukaryotic neurochondrin proteins Neurochondrin induces hydroxyapatite resorptive activity in bone marrow cells resistant to bafilomycin A1, an inhibitor of macrophage- and osteoclast-mediated resorption.

Table A1 (continued)

#	# of exons	length (aa)	Domain Name	Description of the function
20	4	741	FHA	Associated domain (FHA); found in eukaryotic and prokaryotic proteins. In eukaryotes, many FHA domain-containing proteins localize to the nucleus, where they participate in establishing or maintaining cell cycle checkpoints, DNA repair, or transcriptional regulation.
21	1	875	Nup84_Nup100	Core protein 84 / 107 Nup84p forms a complex with five proteins, of which Nup120p, Nup85p, Sec13p, and a Sec13p homologues.
22	8	1062	CA	Repeats domain; involved in Ca2+-mediated cell-cell adhesion; plays a role in cell fate, signaling, proliferation, differentiation, and migration.
23	3	202	MEA1	Enhanced antigen 1 (MEA1). Consists of several mammalian male enhanced antigen 1 (MEA1) proteins. The Mea1 gene is found to be localized in primary and secondary spermatocytes and spermatids. The protein product is highly similar to that of Drosophila CG14341-PB.
c) *C. elegans*				
1	2	102	SapB	Saposin (B) Domains. Present in multiple copies in prosaposin and in pulmonary surfactant-associated protein B. In plant aspartic proteinases, a saposin domain is circularly permuted.
2	3	137	TRS20	Subunit of TRAPP, an ER-Golgi tethering complex Cell motility and secretion.
3	2	139	DUF290	Transthyretin-like family. Similarity to transthyretin. The specific function of this protein is unknown.
4	2	142	DIM1	Mitosis protein DIM1.

Table A1 (continued)

#	# of exons	length (aa)	Domain Name	Description of the function
5	9	578	RasGAP	GTPase-activator protein for Ras-like GTPases. All alpha-helical domain that accelerates the GTPase activity of Ras, thereby "switching" it into an "off" position. Improved domain limits from structure.
6	3	184	-	TetC [Shigella flexneri] [Shigella flexneri], tetracycline resistance protein C.
d) *C. intestinalis*				
1	2	68	GGL	Involved in signal transduction via G-protein-coupled receptors.
2	2	83	zf-CSL	CSL zinc finger. The molecular function of is uncertain
3	1	102		Chaperonin 10 Kd subunit in protein folding ATP binding.
4	3	152	MAPEG family	Membrane associated proteins in Eicosanoid and Glutathione metabolism. Catalyses the synthesis of PGE2 from PGH2 .
5	1	231	PCMT	Protein-L-isoaspartate (D-aspartate) O-methyltransferase activity.
6	5	244	L10	Ribosomal protein.
7	3	249	MIT	Microtubule interacting and trafficking molecule domain.
8	1	286	RplO	Ribosomal protein
9	3	319	Per1	A member of this family has been implemented in protein processing in the endoplasmic reticulum
10	1	407	RNA_pol_I_A49	A49-like RNA polymerase I associated factor. Involved in transcription of ribosomal DNA.
11	5	483	PTB	Phosphotyrosine-binding (PTB) domain
12	1	517	CAP_ED	Effector domain of the CAP family of transcription factors.
13	2	764	eIF3c_N	Eukaryotic translation initiation factor 3 subunit 8 N-terminus. The largest of the mammalian translation initiation factors.

Table A1 (continued)

#	# of exons	length (aa)	Domain Name	Description of the function
e) *C. reinhardtii*				
1	3	56	Ribosomal_S21e	Translation, ribosomal structure and biogenesis.
2	2	64	Nop10p	Nucleolar RNA-binding protein, Nop10p family. Essential for 18S rRNA production and rRNA pseudouridylation.
3	3	89	CSL zinc finger	Zinc binding motif which contains four cysteine residues which chelate zinc. This domain is often found associated with a pfam00226 domain.
4	4	92	Sec61beta family	Component of the Sec61/SecYEG protein secretory system.
5	4	101	Uncharacterized protein family	Not characterized.
6	3	103	ribosomal protein S21	Small ribosomal subunit.
7	3	110	SRP9	Signal recognition particle 9 kDa protein. Pausing of synthesis of ribosome associated nascent polypeptides that have been engaged by the targeting domain of SRP.
8	1	129	ACN9 family	Localized to the mitochondrial inter-membrane space may be a necessary general component of gluconeogenesis.
9	3	130	SCP2	SCP-2 sterol transfer family. Involved in binding sterols.
10	5	134	Ribosomal protein S9/S16	Ribosomal subunit.
11	3	162	SAP	Predicted to be involved in chromosomal organization.
12	4	236	U1-like zinc finger	Zinc ion binding, nucleic acid binding.
f) *D. melanogaster*				
1	1	73	-	TTD-A gene involved in stabilizing of the basal transcription complex TFIIH.

Table A1 (continued)

#	# of exons	length (aa)	Domain Name	Description of the function
2	4	137	Complex1_LYR	Complex 1 protein (LYR family). This family of short proteins includes proteins from the NADH-ubiquinone oxidoreductase complex I, B14 subunit from Cow, and the B22 subunit from human.
3	1	150	Complex1_17_ 2kD	NADH:ubiquinone oxidoreductase 17.2 kD subunit. This family contains the 17.2 kD subunit of complex I and its homologues. The family also contains a second related eukaryotic protein of unknown function..

Table A2. List of newly identified and functionally characterized proteins in 16 fungal genomes.

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
		in query	in target		
chromosome_AM27098_Aspergillus_niger	gene_7	0.97	1.00	gnl|CDD|71315 pfam07876, Dabb, Stress responsive A/B Barrel Domain.	6.00E-13
chromosome_AM270981_Aspergillus_niger	gene_142	0.22	0.96	gnl|CDD|68370 pfam04795, PAPA-1, PAPA-1-like conserved region.	2.00E-07
chromosome_AM270981_Aspergillus_niger	gene_594	0.99	0.93	5' exoribonuclease.	2.00E-47
chromosome_AM270981_Aspergillus_niger	gene_607	0.51	0.59	gnl|CDD|31007 COG0663, PaaY, Carbonic anhydrases/acetyltransferases.	2.00E-11
chromosome_AM270981_Aspergillus_niger	gene_903	0.43	0.95	gnl|CDD|32190 COG2007, RPS8A, Ribosomal protein S8E.	3.00E-18
chromosome_AM270981_Aspergillus_niger	gene_1088	0.81	1.00	gnl|CDD|72040 pfam08615, RNase_H1_sml, Ribonuclease H1 small subunit.	1.00E-29
chromosome_AM270983_Aspergillus_niger	gene_222	0.33	1.00	gnl|CDD|30606 COG0257, RpmJ, Ribosomal protein L36.	3.00E-09
chromosome_AM270983_Aspergillus_niger	gene_725	0.97	1.00	gnl|CDD|31565 COG1374, NIP7, Protein involved in ribosomal biogenesis.	4.00E-39
chromosome_AM270987_Aspergillus_niger	gene_297	0.59	0.80	gnl|CDD|69301 pfam05768, DUF836, Glutaredoxin-like domain (DUF836).	4.00E-07
chromosome_AM270987_Aspergillus_niger	gene_318	0.87	0.94	gnl|CDD|74917 PRK05395, PRK05395, 3-dehydroquinate dehydratase.	7.00E-63
chromosome_AM270987_Aspergillus_niger	gene_476	0.91	1.00	gnl|CDD|34019 COG4297, COG4297, Uncharacterized protein.	2.00E-21
chromosome_AM270988_Aspergillus_niger	gene_464	0.60	0.94	gnl|CDD|71653 pfam08219, TOM13, Outer membrane protein TOM13.	1.00E-16

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
		in query	in target		
chromosome_AM270990_Aspergillus_niger	gene_72	1.00	1.00	gnl|CDD|47816 smart00512, Skp1, Found in Skp1 protein family.	9.00E-12
chromosome_AM270990_Aspergillus_niger	gene_135	0.70	0.87	gnl|CDD|68918 pfam05365, UCR_UQCRX_QCR9, Ubiquinol-cytochrome C reductase.	1.00E-10
chromosome_AM270991_Aspergillus_niger	gene_332	0.95	1.00	gnl|CDD|68209 pfam04628, Sedlin_N, Sedlin, N-terminal conserved region.	3.00E-30
chromosome_AM270991_Aspergillus_niger	gene_642	0.41	1.00	gnl|CDD|71661 pfam08227, DASH_Hsk3, DASH complex subunit Hsk3 like.	3.00E-09
chromosome_AM270993_Aspergillus_niger	gene_198	0.85	1.00	gnl|CDD|33491 COG3695, COG3695, Predicted methylated DNA-protein cysteine.	3.00E-18
chromosome_AM270993_Aspergillus_niger	gene_280	0.83	0.83	gnl|CDD|67778 pfam04178, Got1, Got1-like family.	3.00E-15
chromosome_AM270994_Aspergillus_niger	gene_11	0.52	0.92	gnl|CDD|71474 pfam08038, Tom7, TOM7 family.	3.00E-09
chromosome_AM270994_Aspergillus_niger	gene_20	0.83	1.00	gnl|CDD|72555 pfam09138, Urm1, Urm1 (Ubiquitin related modifier).	2.00E-30
chromosome_AM270996_Aspergillus_niger	gene_175	0.71	0.96	gnl|CDD|66614 pfam02953, zf-Tim10_DDP, Tim10/DDP family zinc finger.	2.00E-10
chromosome_AM270996_Aspergillus_niger	gene_180	0.95	0.91	gnl|CDD|32119 COG1936, COG1936, Predicted nucleotide kinase.	2.00E-35

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_1.1_of_botrytis_cinerea	gene_4	0.95	0.67	gnl|CDD|69275 pfam05742, DUF833, Protein of unknown function (DUF833).	3.00E-26
supercontig_1.26_of_botrytis_cinerea	gene_1	0.75	1.00	gnl|CDD|58528 cd04413, NDPk_I, Nucleoside diphosphate kinase Group I.	2.00E-62
supercontig_1.32_of_botrytis_cinerea	gene_53	0.83	1.00	gnl|CDD|31172 COG0830, UreF, Urease accessory protein UreF.	3.00E-22
supercontig_1.46_of_botrytis_cinerea	gene_1	0.65	0.75	gnl|CDD|29325 cd00250, CAS_like, Clavaminic acid synthetase (CAS) -like.	2.00E-09
supercontig_1.49_of_botrytis_cinerea	gene_12	0.89	1.00	gnl|CDD|33440 COG3642, COG3642, Mn2+-dependent serine/threonine protein kinase.	6.00E-54
supercontig_1.54_of_botrytis_cinerea	gene_8	1.00	1.00	gnl|CDD|79639 pfam00576, Transthyretin, Transthyretin precursor.	1.00E-20
supercontig_1.7_of_botrytis_cinerea	gene_144	0.88	0.87	gnl|CDD|30229 cd02198, YjgH.	2.00E-19
supercontig_1.70_of_botrytis_cinerea	gene_8	0.29	1.00	gnl|CDD|67744 pfam04140, ICMT, Isoprenylcysteine carboxyl methyltransferase.	3.00E-19
supercontig_1.73_of_botrytis_cinerea	gene_13	0.39	0.58	gnl|CDD|30641 COG0293, FtsJ, 23S rRNA methylase.	3.00E-19
supercontig_1.1_of_botrytis_cinerea	gene_147	0.34	1.00	gnl|CDD|71936 pfam08508, DUF1746, Fungal domain of unknown function (DUF1746).	2.00E-28
supercontig_1.100_of_botrytis_cinerea	gene_45	0.31	0.91	gnl|CDD|31133 COG0790, COG0790, FOG: TPR repeat, SEL1 subfamily.	2.00E-15
supercontig_1.103_of_botrytis_cinerea	gene_6	0.69	0.79	gnl|CDD|29325 cd00250, CAS_like, Clavaminic acid synthetase (CAS) -like.	1.00E-10
supercontig_1.103_of_botrytis_cinerea	gene_11	0.27	0.82	gnl|CDD|79643 pfam00583, Acetyltransf_1, Acetyltransferase (GNAT) family.	2.00E-07

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_1.112_of_botrytis_cinerea	gene_36	0.40	0.98	gnl|CDD|70208 pfam06728, PIG-U, GPI transamidase subunit PIG-U.	7.00E-41
supercontig_1.115_of_botrytis_cinerea	gene_14	0.43	1.00	gnl|CDD|71886 pfam08457, Sfi1, Sfi1 spindle body protein.	1.00E-44
supercontig_1.120_of_botrytis_cinerea	gene_5	0.96	0.82	gnl|CDD|70183 pfam06703, SPC25, Microsomal signal peptidase 25 kDa subunit.	3.00E-22
supercontig_1.120_of_botrytis_cinerea	gene_28	0.85	1.00	gnl|CDD|79991 pfam03155, Alg6_Alg8, ALG6, ALG8 glycosyltransferase family.	e-102
supercontig_1.123_of_botrytis_cinerea	gene_6	0.62	0.36	gnl|CDD|48039 cd03444, Thioesterase_II_repeat1.	3.00E-09
supercontig_1.124_of_botrytis_cinerea	gene_3	0.80	1.00	gnl|CDD|71474 pfam08038, Tom7, TOM7 family.	5.00E-10
supercontig_1.132_of_botrytis_cinerea	gene_4	1.00	0.62	gnl|CDD|72075 pfam08650, DASH_Dad4, DASH complex subunit Dad4.	6.00E-06
supercontig_1.132_of_botrytis_cinerea	gene_6	0.82	0.82	gnl|CDD|67778 pfam04178, Got1, Got1-like family.	2.00E-16
supercontig_1.132_of_botrytis_cinerea	gene_17	0.89	0.96	gnl|CDD|29067 cd00148, PROF, Profilin binds actin monomers.	3.00E-20
supercontig_1.164_of_botrytis_cinerea	gene_12	0.98	1.00	gnl|CDD|79555 pfam00300, PGAM, Phosphoglycerate mutase family.	6.00E-20
supercontig_1.165_of_botrytis_cinerea	gene_2	0.37	0.66	gnl|CDD|29021 cd00266, MADS_SRF_like, SRF-like/Type I subfamily of MADS.	4.00E-10
supercontig_1.181_of_botrytis_cinerea	gene_10	0.80	0.99	gnl|CDD|67743 pfam04139, Rad9, Rad9. Rad9 is required for transient cell-cycle.	5.00E-25
supercontig_1.14_of_botrytis_cinerea	gene_15	0.89	0.75	gnl|CDD|30892 COG0546, Gph, Predicted phosphatases.	6.00E-15

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_1.14_of_botrytis_cinerea	gene_17	0.90	0.74	gnl|CDD|30892 COG0546, Gph, Predicted phosphatases.	6.00E-15
supercontig_1.190_of_botrytis_cinerea	gene_1	0.49	0.75	gnl|CDD|47665 smart00338, BRLZ, basic region leucin zipper.	2.00E-08
supercontig_1.16_of_botrytis_cinerea	gene_14	0.92	0.94	gnl|CDD|32448 COG2267, PldB, Lysophospholipase [Lipid metabolism].	1.00E-34
supercontig_1.18_of_botrytis_cinerea	gene_3	0.82	1.00	gnl|CDD|74982 PRK05498, rplF, 50S ribosomal protein L6.	4.00E-36
supercontig_1.2_of_botrytis_cinerea	gene_32	0.78	1.00	gnl|CDD|68096 pfam04511, DER1, Der1-like family.	5.00E-43
supercontig_1.544_of_botrytis_cinerea	gene_1	0.29	0.89	gnl|CDD|48003 smart00736, CADG, Dystroglycan-type cadherin-like domains.	4.00E-06
supercontig_Coccidioides_immitis_RS2.1	gene_469	0.70	0.65	gnl|CDD|74804 PRK05134, PRK05134, 3-demethylubiquinone-9 3-methyltransferase.	8.00E-22
supercontig_Coccidioides_immitis_RS2.1	gene_488	0.95	1.00	gnl|CDD|32448 COG2267, PldB, Lysophospholipase [Lipid metabolism].	1.00E-37
supercontig_Coccidioides_immitis_RS2.1	gene_508	0.60	0.76	gnl|CDD|71997 pfam08571, Yos1, Yos1-like.	3.00E-13
supercontig_Coccidioides_immitis_RS2.1	gene_609	0.17	1.00	gnl|CDD|47943 smart00668, CTLH, C-terminal to LisH motif. Alpha-helical motif.	2.00E-06
supercontig_Coccidioides_immitis_RS2.1	gene_1465	0.88	1.00	gnl|CDD|30620 COG0271, BolA, Stress-induced morphogen (activity unknown).	2.00E-12
supercontig_Coccidioides_immitis_RS2.1	gene_1469	0.67	1.00	gnl|CDD|67724 pfam04119, HSP9_HSP12, Heat shock protein 9 / 12.	1.00E-06
supercontig_Coccidioides_immitis_RS2.1	gene_1842	0.73	0.77	gnl|CDD|69216 pfam05680, ATP-synt_E, ATP synthase E chain.	5.00E-14

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_Coccidioides_immitis_RS2.1	gene_2157	0.79	1.00	gnl|CDD|71474 pfam08038, Tom7, TOM7 family.	2.00E-09
supercontig_Coccidioides_immitis_RS2.2	gene_166	0.74	0.38	gnl|CDD|76259 PRK07764, PRK07764, DNA polymerase III subunits gamma and tau.	6.00E-07
supercontig_Coccidioides_immitis_RS2.2	gene_585	0.31	1.00	gnl|CDD|28942 cd00060, FHA, Forkhead associated domain (FHA).	1.00E-13
supercontig_Coccidioides_immitis_RS2.2	gene_1028	0.68	0.79	gnl|CDD|75062 PRK05649, PRK05649, 4-hydroxybenzoate octaprenyltransferase.	2.00E-47
supercontig_Coccidioides_immitis_RS2.2	gene_1568	0.68	1.00	gnl|CDD|66614 pfam02953, zf-Tim10_DDP, Tim10/DDP family zinc finger.	4.00E-12
supercontig_Coccidioides_immitis_RS2.2	gene_1731	0.60	0.94	gnl|CDD|58650 cd00926, Cyt_c_Oxidase_VIb, Cytochrome c oxidase subunit VIb.	3.00E-26
supercontig_Coccidioides_immitis_RS2.2	gene_1811	0.48	1.00	gnl|CDD|72056 pfam08631, SPO22, Sporulation protein SPO22 like.	2.00E-45
supercontig_Coccidioides_immitis_RS2.2	gene_2059	0.85	1.00	gnl|CDD|31830 COG1644, RPB10, DNA-directed RNA polymerase, subunit N.	3.00E-21
supercontig_Coccidioides_immitis_RS2.3	gene_219	0.21	0.87	gnl|CDD|48627 cd03078, GST_N_Metaxin1_like, GST_N family, Metaxin subfamily.	2.00E-11
supercontig_Coccidioides_immitis_RS2.3	gene_274	0.63	0.89	gnl|CDD|68638 pfam05071, NDUFA12, NADH ubiquinone oxidoreductase subunit NDUFA12.	1.00E-12
supercontig_Coccidioides_immitis_RS2.3	gene_552	0.81	1.00	gnl|CDD|71726 pfam08293, Mit_rib_S27, Mitochondrial ribosomal subunit S27.	7.00E-17

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_Coccidioides_immitis_RS2.3	gene_753	0.44	0.86	gnl|CDD|71653 pfam08219, TOM13, Outer membrane protein TOM13. The TOM13 family.	5.00E-17
supercontig_Coccidioides_immitis_RS2.3	gene_1041	1.00	0.98	gnl|CDD|29166 cd01763, Sumo, Small ubiquitin-related modifier (SUMO) proteins.	1.00E-25
supercontig_Coccidioides_immitis_RS2.3	gene_1157	0.85	0.99	gnl|CDD|65691 pfam01920, Prefoldin_2, Prefoldin subunit.	3.00E-08
supercontig_Coccidioides_immitis_RS2.3	gene_1351	0.90	0.94	gnl|CDD|34731 COG5130, YIP3, Prenylated rab acceptor 1 and related proteins.	2.00E-39
supercontig_Coccidioides_immitis_RS2.4	gene_95	0.90	1.00	gnl|CDD|29719 cd01732, LSm5, The eukaryotic Sm and Sm-like (LSm) proteins.	2.00E-29
supercontig_Coccidioides_immitis_RS2.4	gene_817	0.59	1.00	gnl|CDD|71021 pfam07574, SMC_Nse1, Nse1 non-SMC component of SMC5-6 complex.	2.00E-37
supercontig_Coccidioides_immitis_RS2.4	gene_862	0.30	1.00	gnl|CDD|47687 smart00360, RRM, RNA recognition motif.	3.00E-12
supercontig_Coccidioides_immitis_RS2.5	gene_648	0.58	1.00	gnl|CDD|72775 pfam01900, RNase_P_Rpp14, Rpp14/Pop5 family. tRNA processing enzyme.	5.00E-13
supercontig_2.6_of_Fusarium_oxysporum_f._sp._lycopersici	gene_149	0.39	1.00	gnl|CDD|71661 pfam08227, DASH_Hsk3, DASH complex subunit Hsk3 like.	1.00E-08
supercontig_2.1_of_Fusarium_oxysporum_f._sp._lycopersici	gene_14	0.09	1.00	gnl|CDD|79619 pfam00520, Ion_trans, Ion transport protein.	6.00E-20
supercontig_2.1_of_Fusarium_oxysporum_f._sp._lycopersici	gene_130	0.36	1.00	gnl|CDD|64629 pfam00773, RNB, RNB domain.	8.00E-55
supercontig_2.1_of_Fusarium_oxysporum_f._sp._lycopersici	gene_261	0.65	1.00	gnl|CDD|29029 cd00127, DSPc, Dual specificity phosphatases (DSP);	1.00E-23

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_2.1_of_Fusarium_oxysporum_f._sp._lycopersici	gene_344	0.33	0.40	gnl|CDD|34840 COG5243, HRD1, HRD ubiquitin ligase complex, ER membrane component.	8.00E-11
supercontig_2.1_of_Fusarium_oxysporum_f._sp._lycopersici	gene_605	0.87	0.66	gnl|CDD|29481 cd00687, Terpene_cyclase_nonplant_C1, NonPlant Terpene Cyclases.	1.00E-13
supercontig_2.1_of_Fusarium_oxysporum_f._sp._lycopersici	gene_751	0.58	0.44	gnl|CDD|76259 PRK07764, PRK07764, DNA polymerase III subunits gamma and tau.	1.00E-06
supercontig_2.1_of_Fusarium_oxysporum_f._sp._lycopersici	gene_866	0.95	1.00	gnl|CDD|72346 pfam08927, DUF1909, Domain of unknown function (DUF1909).	1.00E-11
supercontig_2.25_of_Fusarium_oxysporum_f._sp._lycopersici	gene_243	0.14	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	5.00E-68
supercontig_2.26_of_Fusarium_oxysporum_f._sp._lycopersici	gene_218	0.22	0.92	gnl|CDD|67688 pfam04082, Fungal_trans, Fungal specific transcription factor.	1.00E-09
supercontig_2.26_of_Fusarium_oxysporum_f._sp._lycopersici	gene_229	0.66	0.77	gnl|CDD|32959 COG3145, AlkB, Alkylated DNA repair protein.	5.00E-09
supercontig_2.26_of_Fusarium_oxysporum_f._sp._lycopersici	gene_312	0.40	0.96	gnl|CDD|28826 cd02164, PPAT_CoAS.	3.00E-26
supercontig_2.28_of_Fusarium_oxysporum_f._sp._lycopersici	gene_138	1.00	0.44	gnl|CDD|48392 cd02146, NfsA_FRP.	1.00E-18
supercontig_2.28_of_Fusarium_oxysporum_f._sp._lycopersici	gene_148	0.45	0.64	gnl|CDD|69235 pfam05699, hATC, hAT family dimerisation domain.	6.00E-07
supercontig_2.28_of_Fusarium_oxysporum_f._sp._lycopersici	gene_1	0.33	0.71	Integrase mediates.	2.00E-11
supercontig_2.30_of_Fusarium_oxysporum_f._sp._lycopersici	gene_104	1.00	0.44	gnl|CDD|48392 cd02146, NfsA_FRP.	1.00E-18

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_2.30_of_Fusarium_oxysporum_f._sp._lycopersici	gene_114	0.45	0.64	gnl|CDD|69235 pfam05699, hATC, hAT family dimerisation domain.	6.00E-07
supercontig_2.31_of_Fusarium_oxysporum_f._sp._lycopersici	gene_72	0.14	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR: Reverse transcriptases (RTs).	5.00E-68
supercontig_2.31_of_Fusarium_oxysporum_f._sp._lycopersici	gene_73	0.35	0.75	gnl|CDD|64525 pfam00665, rve, Integrase core domain.	4.00E-12
supercontig_2.31_of_Fusarium_oxysporum_f._sp._lycopersici	gene_3	0.74	0.95	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	3.00E-44
supercontig_2.4_of_Fusarium_oxysporum_f._sp._lycopersici	gene_941	0.15	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	1.00E-67
supercontig_2.4_of_Fusarium_oxysporum_f._sp._lycopersici	gene_947	0.86	1.00	gnl|CDD|30876 COG0530, ECM27, Ca2+/Na+ antiporter.	3.00E-20
supercontig_2.36_of_Fusarium_oxysporum_f._sp._lycopersici	gene_1	0.18	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	1.00E-67
supercontig_2.37_of_Fusarium_oxysporum_f._sp._lycopersici	gene_17	0.14	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	5.00E-68
supercontig_2.38_of_Fusarium_oxysporum_f._sp._lycopersici	gene_15	0.14	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	5.00E-68
supercontig_2.43_of_Fusarium_oxysporum_f._sp._lycopersici	gene_30	0.10	0.96	gnl|CDD|66556 pfam02891, zf-MIZ, MIZ/SP-RING zinc finger.	3.00E-16
supercontig_2.43_of_Fusarium_oxysporum_f._sp._lycopersici	gene_45	0.34	0.79	gnl|CDD|79449 pfam00075, RnaseH, RNase H.	2.00E-09
supercontig_2.44_of_Fusarium_oxysporum_f._sp._lycopersici	gene_29	0.14	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	5.00E-68
supercontig_2.45_of_Fusarium_oxysporum_f._sp._lycopersici	gene_45	0.14	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	5.00E-68

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_2.48_of_Fusarium_oxysporum_f._sp._lycopersici	gene_25	0.85	1.00	gnl|CDD|48039 cd03444, Thioesterase_II_repeat1.	1.00E-15
supercontig_2.5_of_Fusarium_oxysporum_f._sp._lycopersici	gene_346	0.65	0.57	gnl|CDD|29142 cd00180, S_TKc, Serine/Threonine protein kinases, catalytic domain.	2.00E-08
supercontig_2.5_of_Fusarium_oxysporum_f._sp._lycopersici	gene_426	0.84	1.00	gnl|CDD|69918 pfam06424, PRP1_N, PRP1 splicing factor, N-terminal.	4.00E-32
supercontig_2.5_of_Fusarium_oxysporum_f._sp._lycopersici	gene_727	0.77	1.00	gnl|CDD|31294 COG1097, RRP4, RNA-binding protein Rrp4 and related proteins.	2.00E-40
supercontig_2.54_of_Fusarium_oxysporum_f._sp._lycopersici	gene_2	0.30	0.94	gnl|CDD|64525 pfam00665, rve, Integrase core domain.	5.00E-17
supercontig_2.6_of_Fusarium_oxysporum_f._sp._lycopersici	gene_337	0.55	0.79	gnl|CDD|69734 pfam06229, FRG1, FRG1-like family.	8.00E-06
supercontig_2.74_of_Fusarium_oxysporum_f._sp._lycopersici	gene_5	0.73	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	1.00E-56
supercontig_2.1_of_Fusarium_oxysporum_f._sp._lycopersici	gene_68	0.95	0.96	gnl|CDD|34788 COG5189, SFP1, Putative transcriptional repressor regulating G2/M.	1.00E-31
supercontig_2.96_of_Fusarium_oxysporum_f._sp._lycopersici	gene_2	0.66	0.56	gnl|CDD|32411 COG2230, Cfa, Cyclopropane fatty acid synthase and related.	2.00E-11
supercontig_2.8_of_Fusarium_oxysporum_f._sp._lycopersici	gene_51	0.69	1.00	gnl|CDD|65608 pfam01826, TIL, Trypsin Inhibitor like cysteine rich domain.	5.00E-06
supercontig_2.8_of_Fusarium_oxysporum_f._sp._lycopersici	gene_219	0.82	0.95	gnl|CDD|79719 pfam00903, Glyoxalase, Glyoxalase/Bleomycin resistance.	2.00E-06
supercontig_2.8_of_Fusarium_oxysporum_f._sp._lycopersici	gene_249	0.68	0.92	gnl|CDD|34729 COG5128, COG5128, Transport protein particle (TRAPP) complex.	2.00E-46

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_2.8_of_Fusarium_oxysporum_f._sp._lycopersici	gene_289	0.47	1.00	gnl|CDD|72017 pfam08592, DUF1772, Domain of unknown function (DUF1772).	1.00E-13
supercontig_2.8_of_Fusarium_oxysporum_f._sp._lycopersici	gene_510	0.90	0.87	gnl|CDD|67779 pfam04179, Init_tRNA_PT, Initiator tRNA phosphoribosyl transferase.	1.00E-85
supercontig_2.9_of_Fusarium_oxysporum_f._sp._lycopersici	gene_4	0.20	0.98	gnl|CDD|64525 pfam00665, rve, Integrase core domain.	2.00E-17
supercontig_2.9_of_Fusarium_oxysporum_f._sp._lycopersici	gene_57	0.31	0.93	gnl|CDD|47948 smart00674, CENPB, Putative DNA-binding domain in centromere.	7.00E-06
supercontig_2.9_of_Fusarium_oxysporum_f._sp._lycopersici	gene_115	0.14	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	5.00E-68
supercontig_2.9_of_Fusarium_oxysporum_f._sp._lycopersici	gene_32	0.95	0.65	gnl|CDD|73491 PRK00102, rnc, ribonuclease III.	4.00E-07
supercontig_2.2_of_Fusarium_oxysporum_f._sp._lycopersici	gene_57	0.78	1.00	gnl|CDD|32694 COG2867, COG2867, Oligoketide cyclase/lipid transport protein.	7.00E-26
supercontig_2.2_of_Fusarium_oxysporum_f._sp._lycopersici	gene_163	0.79	1.00	gnl|CDD|71474 pfam08038, Tom7, TOM7 family. This family consists of TOM7 family.	5.00E-10
supercontig_2.10_of_Fusarium_oxysporum_f._sp._lycopersici	gene_114	0.93	1.00	gnl|CDD|31782 COG1594, RPB9, DNA-directed RNA polymerase, subunit M/Transcription.	7.00E-14
supercontig_2.10_of_Fusarium_oxysporum_f._sp._lycopersici	gene_554	0.14	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	5.00E-68
supercontig_2.11_of_Fusarium_oxysporum_f._sp._lycopersici	gene_127	0.96	1.00	gnl|CDD|32441 COG2260, COG2260, Predicted Zn-ribbon RNA-binding protein.	6.00E-10
supercontig_2.11_of_Fusarium_oxysporum_f._sp._lycopersici	gene_95	1.00	1.00	gnl|CDD|71966 pfam08538, DUF1749, Protein of unknown function (DUF1749).	2.00E-71

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_2.11_of_Fusarium_oxysporum_f._sp._lycopersici	gene_190	0.98	1.00	gnl|CDD|30426 COG0077, PheA, Prephenate dehydratase.	7.00E-71
supercontig_2.12_of_Fusarium_oxysporum_f._sp._lycopersici	gene_283	0.17	0.81	gnl|CDD|28964 cd00083, HLH, Helix-loop-helix domain.	3.00E-07
supercontig_2.12_of_Fusarium_oxysporum_f._sp._lycopersici	gene_400	0.26	1.00	gnl|CDD|48037 cd03442, BFIT_BACH, Brown fat-inducible thioesterase (BFIT).	1.00E-20
supercontig_2.2_of_Fusarium_oxysporum_f._sp._lycopersici	gene_70	0.64	1.00	gnl|CDD|72839 pfam08583, UPF0287, Uncharacterised protein family (UPF0287).	1.00E-22
supercontig_2.2_of_Fusarium_oxysporum_f._sp._lycopersici	gene_363	0.91	1.00	gnl|CDD|29719 cd01732, LSm5, The eukaryotic Sm and Sm-like (LSm) proteins.	3.00E-30
supercontig_2.2_of_Fusarium_oxysporum_f._sp._lycopersici	gene_812	0.94	1.00	gnl|CDD|33183 COG3376, HoxN, High-affinity nickel permease.	2.00E-78
supercontig_2.2_of_Fusarium_oxysporum_f._sp._lycopersici	gene_890	0.35	0.75	gnl|CDD|64525 pfam00665, rve, Integrase core domain.	4.00E-12
supercontig_2.2_of_Fusarium_oxysporum_f._sp._lycopersici	gene_891	0.23	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	3.00E-67
supercontig_2.14_of_Fusarium_oxysporum_f._sp._lycopersici	gene_44	0.14	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	5.00E-68
supercontig_2.14_of_Fusarium_oxysporum_f._sp._lycopersici	gene_159	0.70	0.92	gnl|CDD|34746 COG5145, RAD14, DNA excision repair protein.	5.00E-50
supercontig_2.15_of_Fusarium_oxysporum_f._sp._lycopersici	gene_7	0.78	0.62	gnl|CDD|74518 PRK03983, PRK03983, exosome complex exonuclease Rrp41.	3.00E-11
supercontig_2.15_of_Fusarium_oxysporum_f._sp._lycopersici	gene_100	0.62	0.96	gnl|CDD|69491 pfam05971, Methyltransf_10, Protein of unknown function (DUF890).	2.00E-53
supercontig_2.15_of_Fusarium_oxysporum_f._sp._lycopersici	gene_198	0.59	0.55	gnl|CDD|30614 COG0265, DegQ, Trypsin-like serine proteases.	3.00E-07

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_2.15_of_Fusarium_oxysporum_f._sp._lycopersici	gene_289	0.61	1.00	gnl|CDD|47935 smart00659, RPOLCX, RNA polymerase subunit CX.	2.00E-08
supercontig_2.15_of_Fusarium_oxysporum_f._sp._lycopersici	gene_311	0.55	0.49	gnl|CDD|66174 pfam02458, Transferase, Transferase family.	3.00E-15
supercontig_2.17_of_Fusarium_oxysporum_f._sp._lycopersici	gene_54	0.97	1.00	gnl|CDD|72082 pfam08657, DASH_Spc34, DASH complex subunit Spc34.	5.00E-54
supercontig_2.17_of_Fusarium_oxysporum_f._sp._lycopersici	gene_56	0.12	1.00	gnl|CDD|29102 cd00162, RING, RING-finger (Really Interesting New Gene) domain.	3.00E-06
supercontig_2.17_of_Fusarium_oxysporum_f._sp._lycopersici	gene_291	0.54	0.62	gnl|CDD|29841 cd00632, Prefoldin_beta, Prefoldin beta.	6.00E-06
supercontig_2.18_of_Fusarium_oxysporum_f._sp._lycopersici	gene_29	0.49	0.50	gnl|CDD|79444 pfam00069, Pkinase, Protein kinase domain.	5.00E-09
supercontig_2.19_of_Fusarium_oxysporum_f._sp._lycopersici	gene_350	0.53	0.73	gnl|CDD|69043 pfam05498, RALF, Rapid ALkalinization Factor (RALF).	3.00E-13
supercontig_2.20_of_Fusarium_oxysporum_f._sp._lycopersici	gene_180	0.76	0.55	gnl|CDD|69212 pfam05676, NDUF_B7, NADH-ubiquinone oxidoreductase B18 subunit.	3.00E-11
supercontig_2.20_of_Fusarium_oxysporum_f._sp._lycopersici	gene_251	0.23	1.00	gnl|CDD|48011 smart00744, RINGv, The RING-variant domain is a C4HC3 zinc-finger.	8.00E-07
supercontig_2.21_of_Fusarium_oxysporum_f._sp._lycopersici	gene_11	0.49	0.50	gnl|CDD|79444 pfam00069, Pkinase, Protein kinase domain.	5.00E-09
supercontig_2.3_of_Fusarium_oxysporum_f._sp._lycopersici	gene_35	0.89	0.93	gnl|CDD|66174 pfam02458, Transferase, Transferase family.	4.00E-15
supercontig_2.3_of_Fusarium_oxysporum_f._sp._lycopersici	gene_98	0.92	0.98	gnl|CDD|31089 COG0746, MobA, Molybdopterin-guanine dinucleotide biosynthesis.	4.00E-14

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_2.3_of_Fusarium_oxysporum_f._sp._lycopersici	gene_422	0.64	0.43	gnl|CDD|30669 COG0321, LipB, Lipoate-protein ligase B [Coenzyme metabolism].	2.00E-14
supercontig_2.3_of_Fusarium_oxysporum_f._sp._lycopersici	gene_442	0.97	1.00	gnl|CDD|47853 smart00552, ADEAMc, tRNA-specific and double-stranded RNA adenosine.	1.00E-47
supercontig_2.3_of_Fusarium_oxysporum_f._sp._lycopersici	gene_455	0.50	1.00	gnl|CDD|71661 pfam08227, DASH_Hsk3, DASH complex subunit Hsk3 like.	2.00E-08
supercontig_2.3_of_Fusarium_oxysporum_f._sp._lycopersici	gene_515	0.96	1.00	gnl|CDD|69400 pfam05871, ESCRT-II, ESCRT-II complex subunit.	9.00E-50
supercontig_2.3_of_Fusarium_oxysporum_f._sp._lycopersici	gene_654	0.74	1.00	gnl|CDD|31778 COG1590, COG1590, Uncharacterized conserved protein.	6.00E-24
supercontig_2.22_of_Fusarium_oxysporum_f._sp._lycopersici	gene_49	0.14	1.00	gnl|CDD|73154 cd01647, RT_LTR, RT_LTR.	5.00E-68
supercontig_2.23_of_Fusarium_oxysporum_f._sp._lycopersici	gene_241	0.93	0.59	gnl|CDD|30949 COG0604, Qor, NADPH:quinone reductase and related Zn-dependent.	1.00E-17
supercontig_2.23_of_Fusarium_oxysporum_f._sp._lycopersici	gene_336	0.51	0.93	gnl|CDD|29261 cd00204, ANK, ankyrin repeats.	3.00E-08
supercontig_2.24_of_Fusarium_oxysporum_f._sp._lycopersici	gene_60	0.59	1.00	gnl|CDD|47626 smart00298, CHROMO, Chromatin organization modifier domain.	6.00E-06
Sclerotinia_sclerotiorum_supercontig_1.1	gene_495	0.70	0.90	gnl|CDD|68638 pfam05071, NDUFA12, NADH ubiquinone oxidoreductase subunit NDUFA12.	5.00E-12
Sclerotinia_sclerotiorum_supercontig_1.1	gene_748	0.63	0.94	gnl|CDD|47654 smart00326, SH3, Src homology 3 domains; Src homology 3 (SH3).	1.00E-09
Sclerotinia_sclerotiorum_supercontig_1.10	gene_123	0.42	0.80	gnl|CDD|30683 COG0335, RplS, Ribosomal protein L19.	9.00E-10

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
Sclerotinia_sclerotiorum_supercontig_1.10	gene_132	0.59	1.00	gnl|CDD|69040 pfam05495, zf-CHY, CHY zinc finger.	2.00E-09
Sclerotinia_sclerotiorum_supercontig_1.11	gene_143	0.72	1.00	gnl|CDD|72881 cd02885, IPP_Isomerase, Isopentenyl diphosphate (IPP).	2.00E-60
Sclerotinia_sclerotiorum_supercontig_1.11	gene_375	0.63	0.87	gnl|CDD|75796 PRK06849, PRK06849, hypothetical protein.	2.00E-35
Sclerotinia_sclerotiorum_supercontig_1.11	gene_399	0.96	1.00	gnl|CDD|34679 COG5075, COG5075, Uncharacterized conserved protein.	9.00E-62
Sclerotinia_sclerotiorum_supercontig_1.12	gene_320	0.57	0.97	gnl|CDD|71474 pfam08038, Tom7, TOM7 family.	2.00E-09
Sclerotinia_sclerotiorum_supercontig_1.14	gene_110	0.62	1.00	gnl|CDD|31588 COG1398, OLE1, Fatty-acid desaturase [Lipid metabolism].	3.00E-76
Sclerotinia_sclerotiorum_supercontig_1.18	gene_98	0.65	1.00	gnl|CDD|29952 cd00959, DeoC, 2-deoxyribose-5-phosphate aldolase (DERA).	1.00E-40
Sclerotinia_sclerotiorum_supercontig_1.2	gene_526	0.34	1.00	gnl|CDD|29697 cd00593, RIBOc, RIBOc. Ribonuclease III C terminal domain.	4.00E-21
Sclerotinia_sclerotiorum_supercontig_1.20	gene_57	0.61	0.34	gnl|CDD|65423 pfam01624, MutS_I, MutS domain I.	2.00E-08
Sclerotinia_sclerotiorum_supercontig_1.22	gene_154	0.58	0.40	gnl|CDD|69613 pfam06101, DUF946, Plant protein of unknown function (DUF946).	6.00E-10
Sclerotinia_sclerotiorum_supercontig_1.26	gene_15	0.29	1.00	gnl|CDD|67744 pfam04140, ICMT, Isoprenylcysteine carboxyl methyltransferase	2.00E-19
Sclerotinia_sclerotiorum_supercontig_1.31	gene_24	0.66	1.00	gnl|CDD|67469 pfam03856, SUN, Beta-glucosidase (SUN family).	4.00E-53
Sclerotinia_sclerotiorum_supercontig_1.5	gene_309	0.40	0.98	gnl|CDD|70208 pfam06728, PIG-U, GPI transamidase subunit PIG-U.	1.00E-40

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
Sclerotinia_sclerotiorum_supercontig_1.8	gene_263	0.79	1.00	gnl|CDD|69301 pfam05768, DUF836, Glutaredoxin-like domain (DUF836).	5.00E-06
Phaeosphaeria_nodorum_supercontig_1.33	gene_13	0.82	1.00	gnl|CDD|80120 pfam05721, PhyH, Phytanoyl-CoA dioxygenase (PhyH).	4.00E-28
Phaeosphaeria_nodorum_supercontig_1.10	gene_5	0.74	1.00	gnl|CDD|29709 cd01722, Sm_F, The eukaryotic Sm and Sm-like (LSm) proteins.	6.00E-25
Phaeosphaeria_nodorum_supercontig_1.6	gene_132	0.64	0.80	gnl|CDD|71997 pfam08571, Yos1, Yos1-like.	1.00E-16
Phaeosphaeria_nodorum_supercontig_1.7	gene_201	0.63	0.85	gnl|CDD|68195 pfam04614, Pex19, Pex19 protein family.	9.00E-20
supercontig_3.1_of_Fusarium_graminearum	gene_131	0.40	1.00	gnl|CDD|29261 cd00204, ANK, ankyrin repeats.	7.00E-10
supercontig_3.1_of_Fusarium_graminearum	gene_321	0.35	0.31	gnl|CDD|79137 PRK12678, PRK12678, transcription termination factor Rho.	1.00E-07
supercontig_3.4_of_Fusarium_graminearum	gene_334	0.32	0.37	gnl|CDD|77212 PRK09510, tolA, cell envelope integrity inner membrane protein.	9.00E-06
supercontig_3.4_of_Fusarium_graminearum	gene_698	0.97	1.00	gnl|CDD|72839 pfam08583, UPF0287, Uncharacterised protein family (UPF0287).	1.00E-17
supercontig_3.4_of_Fusarium_graminearum	gene_756	0.45	1.00	gnl|CDD|65057 pfam01230, HIT, HIT domain.	6.00E-07
supercontig_3.4_of_Fusarium_graminearum	gene_852	0.29	0.98	gnl|CDD|65933 pfam02194, PXA, PXA domain. This domain is associated with PX.	1.00E-11
supercontig_3.4_of_Fusarium_graminearum	gene_904	0.62	0.92	gnl|CDD|79664 pfam00639, Rotamase, PPIC-type PPIASE domain.	5.00E-11
supercontig_3.4_of_Fusarium_graminearum	gene_1192	0.65	0.77	gnl|CDD|32959 COG3145, AlkB, Alkylated DNA repair protein.	8.00E-09

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_3.5_of_Fusarium_graminearum	gene_343	0.23	1.00	gnl|CDD|70445 pfam06978, POP1, Ribonucleases P/MRP protein subunit POP1.	2.00E-16
supercontig_3.6_of_Fusarium_graminearum	gene_214	0.23	0.95	gnl|CDD|79643 pfam00583, Acetyltransf_1, Acetyltransferase (GNAT) family.	4.00E-07
supercontig_3.6_of_Fusarium_graminearum	gene_367	0.22	0.98	gnl|CDD|79450 pfam00076, RRM_1, RNA recognition motif. (a.k.a. RRM, RBD, or RNP	3.00E-10
supercontig_3.1_of_Fusarium_graminearum	gene_733	0.82	1.00	gnl|CDD|69301 pfam05768, DUF836, Glutaredoxin-like domain (DUF836).	2.00E-09
supercontig_3.1_of_Fusarium_graminearum	gene_1030	0.79	0.56	gnl|CDD|34793 COG5194, APC11, Component of SCF ubiquitin ligase .	2.00E-08
supercontig_3.6_of_Fusarium_graminearum	gene_347	0.84	1.00	gnl|CDD|31782 COG1594, RPB9, DNA-directed RNA polymerase, subunit M/Transcription.	1.00E-13
supercontig_3.7_of_Fusarium_graminearum	gene_99	0.79	1.00	gnl|CDD|31294 COG1097, RRP4, RNA-binding protein Rrp4 and related proteins.	1.00E-40
supercontig_3.1_of_Fusarium_graminearum	gene_37	0.83	0.95	gnl|CDD|30036 cd01293, Bact_CD, Bacterial cytosine deaminase.	1.00E-48
supercontig_3.1_of_Fusarium_graminearum	gene_85	0.67	0.76	gnl|CDD|72809 pfam05493, ATP_synt_H, ATP synthase subunit H.	2.00E-06
supercontig_3.1_of_Fusarium_graminearum	gene_155	0.31	1.00	gnl|CDD|29261 cd00204, ANK, ankyrin repeats.	4.00E-12
supercontig_3.2_of_Fusarium_graminearum	gene_379	0.86	1.00	gnl|CDD|72162 pfam08738, Gon7, Gon7 family.	1.00E-09
supercontig_3.3_of_Fusarium_graminearum	gene_654	0.70	1.00	gnl|CDD|72775 pfam01900, RNase_P_Rpp14, Rpp14/Pop5 family.	3.00E-14

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
supercontig_3.3_of_Fusarium_graminearum	gene_670	0.59	1.00	gnl|CDD|71719 pfam08286, Spc24, Spc24 subunit of Ndc80.	5.00E-26
supercontig_3.3_of_Fusarium_graminearum	gene_50	0.78	0.93	gnl|CDD|28942 cd00060, FHA, Forkhead associated domain (FHA).	7.00E-10
supercontig_3.3_of_Fusarium_graminearum	gene_239	0.40	1.00	gnl|CDD|71675 pfam08241, Methyltransf_11, Methyltransferase domain.	7.00E-11
Magnaporthe_grisea_70-15_supercontig_5.134	gene_29	0.47	1.00	gnl|CDD|71198 pfam07757, DUF1613, Protein of unknown function (DUF1613).	4.00E-76
Magnaporthe_grisea_70-15_supercontig_5.134	gene_58	0.42	0.97	gnl|CDD|79702 pfam00787, PX, PX domain. PX domains bind to phosphoinositides.	2.00E-17
Magnaporthe_grisea_70-15_supercontig_5.134	gene_167	0.97	0.82	gnl|CDD|30800 COG0451, WcaG, Nucleoside-diphosphate-sugar epimerases .	4.00E-07
Magnaporthe_grisea_70-15_supercontig_5.134	gene_168	0.50	0.76	gnl|CDD|70740 pfam07287, DUF1446, Protein of unknown function (DUF1446).	2.00E-52
Magnaporthe_grisea_70-15_supercontig_5.134	gene_186	0.37	0.82	gnl|CDD|71939 pfam08511, COQ9, COQ9.	2.00E-20
Magnaporthe_grisea_70-15_supercontig_5.178	gene_119	0.65	1.00	gnl|CDD|72775 pfam01900, RNase_P_Rpp14, Rpp14/Pop5 family.	8.00E-15
Magnaporthe_grisea_70-15_supercontig_5.186	gene_45	0.56	0.36	gnl|CDD|76117 PRK07479, PRK07479, 3-ketoacyl-(acyl-carrier-protein).	1.00E-11
Magnaporthe_grisea_70-15_supercontig_5.187	gene_111	0.87	0.99	gnl|CDD|31215 COG1011, COG1011, Predicted hydrolase (HAD superfamily).	2.00E-11
Magnaporthe_grisea_70-15_supercontig_5.187	gene_85	0.72	0.56	gnl|CDD|32715 COG2890, HemK, Methylase of polypeptide chain release factors.	2.00E-09
Magnaporthe_grisea_70-15_supercontig_5.190	gene_81	0.90	0.88	gnl|CDD|35095 COG5536, BET4, Protein prenyltransferase, alpha subunit	1.00E-28

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
Magnaporthe_grisea_70-15_supercontig_5.190	gene_266	0.97	0.93	gnl|CDD|30426 COG0077, PheA, Prephenate dehydratase.	4.00E-58
Magnaporthe_grisea_70-15_supercontig_5.191	gene_38	0.76	1.00	gnl|CDD|64803 pfam00955, HCO3_cotransp, HCO3- transporter family.	1.00E-52
Magnaporthe_grisea_70-15_supercontig_5.194	gene_6	0.63	1.00	gnl|CDD|31137 COG0794, GutQ, Predicted sugar phosphate isomerase.	3.00E-24
Magnaporthe_grisea_70-15_supercontig_5.194	gene_68	0.95	0.75	gnl|CDD|48573 cd03024, DsbA_FrnE, DsbA family, FrnE subfamily.	4.00E-18
Magnaporthe_grisea_70-15_supercontig_5.194	gene_68	0.63	0.97	gnl|CDD|66296 pfam02594, DUF167, Uncharacterised ACR, YggU family COG1872.	2.00E-09
Magnaporthe_grisea_70-15_supercontig_5.194	gene_616	0.57	0.87	gnl|CDD|74772 PRK05014, hscB, co-chaperone HscB.	3.00E-11
Magnaporthe_grisea_70-15_supercontig_5.194	gene_54	0.56	0.97	gnl|CDD|29621 cd00520, RRF, Ribosome recycling factor (RRF).	1.00E-08
Magnaporthe_grisea_70-15_supercontig_5.195	gene_12	0.90	1.00	gnl|CDD|29648 cd00563, Dtyr_deacylase, D-Tyrosyl-tRNAtyr deacylases.	6.00E-46
Magnaporthe_grisea_70-15_supercontig_5.195	gene_75	0.83	0.90	gnl|CDD|75144 PRK05766, rps14P, 30S ribosomal protein S14P.	3.00E-09
Magnaporthe_grisea_70-15_supercontig_5.195	gene_144	0.59	0.81	gnl|CDD|72817 pfam06331, Tbf5, Transcription factor TFIIH complex subunit Tfb5.	6.00E-10
Magnaporthe_grisea_70-15_supercontig_5.196	gene_20	1.00	1.00	gnl|CDD|30839 COG0493, GltD, NADPH-dependent glutamate synthase beta chain.	8.00E-32
Magnaporthe_grisea_70-15_supercontig_5.196	gene_85	0.95	1.00	gnl|CDD|72839 pfam08583, UPF0287, Uncharacterised protein family (UPF0287).	1.00E-16

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
Magnaporthe_grisea_70-15_supercontig_5.196	gene_86	0.83	0.85	gnl|CDD|35107 COG5548, COG5548, Small integral membrane protein.	9.00E-11
Magnaporthe_grisea_70-15_supercontig_5.196	gene_88	0.57	0.41	gnl|CDD|79487 pfam00144, Beta-lactamase, Beta-lactamase.	2.00E-11
Magnaporthe_grisea_70-15_supercontig_5.196	gene_4	0.95	0.87	gnl|CDD|68035 pfam04446, Thg1, tRNAHis guanylyltransferase.	3.00E-53
Magnaporthe_grisea_70-15_supercontig_5.196	gene_168	0.56	0.47	gnl|CDD|29261 cd00204, ANK, ankyrin repeats; ankyrin repeats mediate.	1.00E-06
Magnaporthe_grisea_70-15_supercontig_5.196	gene_188	0.57	1.00	gnl|CDD|69428 pfam05903, DUF862, PPPDE putative peptidase domain.	2.00E-33
AACW02000010_|_CONTIG_10__|_Rhizopus_oryzae_supercontig_3.1_|_[2400983-2598834]_|_197852_nt_	gene_22	0.68	1.00	gnl|CDD|30376 COG0026, PurK, Phosphoribosylaminoimidazole carboxylase.	e-113
AACW02000102_|_CONTIG_102__|_Rhizopus_oryzae_supercontig_3.5_|_[154653-488518]_|_333866_nt_	gene_75	0.73	0.94	gnl|CDD|72061 pfam08636, Pkr1, ER protein Pkr1.	2.00E-06
AACW02000011_|_CONTIG_11__|_Rhizopus_oryzae_supercontig_3.1_|_[2598935-3032215]_|_433281_nt_	gene_16	0.90	1.00	gnl|CDD|30181 cd01994, Alpha_ANH_like_IV.	5.00E-29
AACW02000011_|_CONTIG_11__|_Rhizopus_oryzae_supercontig_3.1_|_[2598935-3032215]_|_433281_nt_	gene_47	0.26	0.95	gnl|CDD|47750 smart00443, G_patch, glycine rich nucleic binding domain.	1.00E-06

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000011_|_CONTIG_11__|_Rhizopus_oryzae_supercontig_3.1_|_[2598935-3032215]_|_433281_nt_	gene_111	0.91	1.00	gnl|CDD|29719 cd01732, LSm5, The eukaryotic Sm and Sm-like (LSm) proteins.	9.00E-30
AACW02000113_|_CONTIG_113__|_Rhizopus_oryzae_supercontig_3.5_|_[1292705-1448586]_|_155882_nt_	gene_44	0.97	0.96	gnl|CDD|34696 COG5093, COG5093, Uncharacterized conserved protein.	1.00E-36
AACW02000116_|_CONTIG_116__|_Rhizopus_oryzae_supercontig_3.5_|_[1648264-1804041]_|_155778_nt_	gene_53	0.55	0.82	gnl|CDD|79643 pfam00583, Acetyltransf_1, Acetyltransferase (GNAT) family.	2.00E-07
AACW02000117_|_CONTIG_117__|_Rhizopus_oryzae_supercontig_3.5_|_[1804142-1926323]_|_122182_nt_	gene_12	0.38	0.98	gnl|CDD|68297 pfam04719, TAFII28, hTAFII28-like protein conserved region.	9.00E-27
AACW02000121_|_CONTIG_121__|_Rhizopus_oryzae_supercontig_3.5_|_[2206362-2288780]_|_82419_nt_	gene_7	0.97	1.00	gnl|CDD|34827 COG5230, COG5230, Uncharacterized conserved protein.	9.00E-31
AACW02000121_|_CONTIG_121__|_Rhizopus_oryzae_supercontig_3.5_|_[2206362-2288780]_|_82419_nt_	gene_21	0.16	0.95	gnl|CDD|32274 COG2091, Sfp, Phosphopantetheinyl transferase.	2.00E-24

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000143_|_CONTIG_143__|_Rhizopus_oryzae_supercontig_3.6_|_[362262-593047]_|_230786_nt_	gene_54	0.86	1.00	gnl|CDD|79999 pfam03372, Exo_endo_phos, Endonuclease/Exonuclease/phosphatase.	5.00E-14
AACW02000143_|_CONTIG_143__|_Rhizopus_oryzae_supercontig_3.6_|_[362262-593047]_|_230786_nt_	gene_57	0.78	0.99	gnl|CDD|71649 pfam08215, DUF1715, Eukaryotic domain of unknown function.	8.00E-15
AACW02000144_|_CONTIG_144__|_Rhizopus_oryzae_supercontig_3.6_|_[593148-906311]_|_313164_nt_	gene_49	0.96	1.00	gnl|CDD|33417 COG3618, COG3618, Predicted metal-dependent hydrolase.	2.00E-38
AACW02000157_|_CONTIG_157__|_Rhizopus_oryzae_supercontig_3.6_|_[2522494-2859520]_|_337027_nt_	gene_16	0.29	0.92	gnl|CDD|29006 cd02396, PCBP_like_KH, K homology RNA-binding domain, PCBP_like.	2.00E-06
AACW02000166_|_CONTIG_166__|_Rhizopus_oryzae_supercontig_3.7_|_[471419-758678]_|_287260_nt_	gene_72	0.96	1.00	gnl|CDD|72167 pfam08743, Nse4, Nse4. Nse4 is a component of the Smc5/6 DNA repair.	2.00E-53
AACW02000173_|_CONTIG_173__|_Rhizopus_oryzae_supercontig_3.7_|_[1109583-1767534]_|_657952_nt_	gene_238	0.86	0.98	gnl|CDD|33461 COG3663, Mug, G:T/U mismatch-specific DNA glycosylase.	2.00E-19

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000194_|_CONTIG_194__|_Rhizopus_oryzae_supercontig_3.8_|_[1456944-1811534]_|_354591_nt_	gene_96	0.87	1.00	gnl|CDD|32648 COG2820, Udp, Uridine phosphorylase.	6.00E-19
AACW02000198_|_CONTIG_198__|_Rhizopus_oryzae_supercontig_3.8_|_[2139510-2362744]_|_223235_nt_	gene_27	0.60	0.34	gnl|CDD|79137 PRK12678, PRK12678, transcription termination factor Rho.	3.00E-09
AACW02000210_|_CONTIG_210__|_Rhizopus_oryzae_supercontig_3.9_|_[1865000-2296105]_|_431106_nt_	gene_22	0.95	0.98	gnl|CDD|78584 PRK11587, PRK11587, putative phosphatase.	2.00E-36
AACW02000216_|_CONTIG_216__|_Rhizopus_oryzae_supercontig_3.10_|_[335899-760890]_|_424992_nt_	gene_23	0.39	0.81	gnl|CDD|29705 cd01718, Sm_E, The eukaryotic Sm and Sm-like (LSm) proteins.	2.00E-25
AACW02000216_|_CONTIG_216__|_Rhizopus_oryzae_supercontig_3.10_|_[335899-760890]_|_424992_nt_	gene_151	0.84	0.88	gnl|CDD|31007 COG0663, PaaY, Carbonic anhydrases/acetyltransferases, isoleucine.	7.00E-16
AACW02000219_|_CONTIG_219__|_Rhizopus_oryzae_supercontig_3.10_|_[1035221-1305776]_|_270556_nt_	gene_68	0.28	0.92	gnl|CDD|67722 pfam04117, Mpv17_PMP22, Mpv17 / PMP22 family.	7.00E-12

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000219_|_CONTIG_219__|_Rhizopus_oryzae_supercontig_3.10_|_[1035221-1305776]_|_270556_nt_	gene_72	0.72	0.76	gnl|CDD|68849 pfam05292, MCD, Malonyl-CoA decarboxylase (MCD).	2.00E-32
AACW02000228_|_CONTIG_228__|_Rhizopus_oryzae_supercontig_3.11_|_[1-736642]_|_736642_nt_	gene_154	0.53	0.76	gnl|CDD|34689 COG5085, COG5085, Predicted membrane protein.	3.00E-19
AACW02000023_|_CONTIG_23__|_Rhizopus_oryzae_supercontig_3.1_|_[3465973-3499457]_|_33485_nt_	gene_5	0.57	1.00	gnl|CDD|69614 pfam06102, DUF947, Domain of unknown function (DUF947).	2.00E-11
AACW02000233_|_CONTIG_233__|_Rhizopus_oryzae_supercontig_3.11_|_[789332-930552]_|_141221_nt_	gene_33	0.41	0.82	gnl|CDD|30935 COG0590, CumB, Cytosine/adenosine deaminases.	1.00E-17
AACW02000235_|_CONTIG_235__|_Rhizopus_oryzae_supercontig_3.11_|_[934731-1439178]_|_504448_nt_	gene_92	0.98	0.91	gnl|CDD|34622 COG5017, COG5017, Uncharacterized conserved protein.	1.00E-13
AACW02000235_|_CONTIG_235__|_Rhizopus_oryzae_supercontig_3.11_|_[934731-1439178]_|_504448_nt_	gene_118	0.95	1.00	gnl|CDD|29161 cd00754, MoaD, MoaD family.	6.00E-14

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000235_|_CONTIG_235__|_Rhizopus_oryzae_supercontig_3.11_|_[934731-1439178]_|_504448_nt_	gene_147	0.76	0.95	gnl|CDD|34813 COG5216, COG5216, Uncharacterized conserved protein.	3.00E-18
AACW02000235_|_CONTIG_235__|_Rhizopus_oryzae_supercontig_3.11_|_[934731-1439178]_|_504448_nt_	gene_171	0.17	1.00	gnl|CDD|72053 pfam08628, Nexin_C, Sorting nexin C terminal.	2.00E-07
AACW02000239_|_CONTIG_239__|_Rhizopus_oryzae_supercontig_3.11_|_[1471025-1585522]_|_114498_nt_	gene_25	0.65	1.00	gnl|CDD|72882 cd03424, ADPRase_NUDT5, ADP-ribose pyrophosphatase (ADPRase)	5.00E-15
AACW02000250_|_CONTIG_250__|_Rhizopus_oryzae_supercontig_3.12_|_[473879-1120864]_|_646986_nt_	gene_8	0.59	0.40	gnl|CDD|68494 pfam04922, DIE2_ALG10, DIE2/ALG10 family.	3.00E-26
AACW02000250_|_CONTIG_250__|_Rhizopus_oryzae_supercontig_3.12_|_[473879-1120864]_|_646986_nt_	gene_80	1.00	1.00	gnl|CDD|67275 pfam03647, TMEM14, Transmembrane proteins 14C.	4.00E-11
AACW02000250_|_CONTIG_250__|_Rhizopus_oryzae_supercontig_3.12_|_[473879-1120864]_|_646986_nt_	gene_142	0.77	0.79	gnl|CDD|70129 pfam06645, SPC12, Microsomal signal peptidase 12 kDa subunit	1.00E-11

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000250_|_CONTIG_250__|_Rhizopus_oryzae_supercontig_3.12_|_[473879-1120864]_|_646986_nt_	gene_218	0.97	1.00	gnl|CDD|68156 pfam04573, SPC22, Signal peptidase subunit.	4.00E-28
AACW02000253_|_CONTIG_253__|_Rhizopus_oryzae_supercontig_3.12_|_[1241831-1398439]_|_156609_nt_	gene_25	0.71	1.00	gnl|CDD|47999 smart00731, SprT, SprT homologues.	9.00E-21
AACW02000257_|_CONTIG_257__|_Rhizopus_oryzae_supercontig_3.13_|_[7284-135576]_|_128293_nt_	gene_32	0.88	0.90	gnl|CDD|68801 pfam05241, EBP, Emopamil binding protein.	1.00E-26
AACW02000268_|_CONTIG_268__|_Rhizopus_oryzae_supercontig_3.14_|_[586241-956944]_|_370704_nt_	gene_6	0.55	1.00	gnl|CDD|79643 pfam00583, Acetyltransf_1, Acetyltransferase (GNAT) family.	1.00E-14
AACW02000268_|_CONTIG_268__|_Rhizopus_oryzae_supercontig_3.14_|_[586241-956944]_|_370704_nt_	gene_130	0.56	0.58	gnl|CDD|35095 COG5536, BET4, Protein prenyltransferase, alpha subunit.	1.00E-08
AACW02000274_|_CONTIG_274__|_Rhizopus_oryzae_supercontig_3.15_|_[1-410886]_|_410886_nt_	gene_112	0.83	0.82	gnl|CDD|70180 pfam06699, PIG-F, Phospho-ethanolamine N-methyltransferase.	1.00E-18

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000275_|_CONTIG_275__|_Rhizopus_oryzae_supercontig_3.15_|_[415307-518953]_|_103647_nt_	gene_4	0.46	0.56	gnl|CDD|30003 cd01393, recA_like, RecA is a bacterial enzyme.	1.00E-07
AACW02000279_|_CONTIG_279__|_Rhizopus_oryzae_supercontig_3.15_|_[570083-942547]_|_372465_nt_	gene_3	0.93	0.94	gnl|CDD|69821 pfam06320, GCN5L1, GCN5-like protein 1 (GCN5L1).	1.00E-20
AACW02000279_|_CONTIG_279__|_Rhizopus_oryzae_supercontig_3.15_|_[570083-942547]_|_372465_nt_	gene_41	0.35	0.80	gnl|CDD|64946 pfam01111, CKS, Cyclin-dependent kinase regulatory subunit.	2.00E-17
AACW02000288_|_CONTIG_288__|_Rhizopus_oryzae_supercontig_3.16_|_[333133-673130]_|_339998_nt_	gene_70	0.44	0.36	gnl|CDD|34689 COG5085, COG5085, Predicted membrane protein.	2.00E-12
AACW02000293_|_CONTIG_293__|_Rhizopus_oryzae_supercontig_3.17_|_[1-476381]_|_476381_nt_	gene_49	0.81	0.65	gnl|CDD|68305 pfam04727, ELMO_CED12, ELMO/CED-12 family.	2.00E-26
AACW02000293_|_CONTIG_293__|_Rhizopus_oryzae_supercontig_3.17_|_[1-476381]_|_476381_nt_	gene_62	0.22	0.97	gnl|CDD|65389 pfam01585, G-patch, G-patch domain.	1.00E-06

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000293_|_CONTIG_293__|_Rhizopus_oryzae_supercontig_3.17_|_[1-476381]_|_476381_nt_	gene_168	0.81	1.00	gnl|CDD|29769 cd01042, DMQH, Demethoxyubiquinone hydroxylases (DMQH).	5.00E-55
AACW02000003_|_CONTIG_3__|_Rhizopus_oryzae_supercontig_3.1_|_[346323-756733]_|_410411_nt_	gene_127	0.64	0.62	gnl|CDD|71935 pfam08507, COPI_assoc, COPI associated protein.	4.00E-07
AACW02000311_|_CONTIG_311__|_Rhizopus_oryzae_supercontig_3.20_|_[152419-490205]_|_337787_nt_	gene_38	0.85	0.64	gnl|CDD|58647 cd00756, MoaE, MoaE family.	3.00E-21
AACW02000313_|_CONTIG_313__|_Rhizopus_oryzae_supercontig_3.21_|_[1-398826]_|_398826_nt_	gene_66	0.95	1.00	gnl|CDD|66475 pfam02792, Mago_nashi.	4.00E-59
AACW02000327_|_CONTIG_327__|_Rhizopus_oryzae_supercontig_3.25_|_[83094-182798]_|_99705_nt_	gene_29	1.00	1.00	gnl|CDD|30755 COG0406, GpmB, Fructose-2,6-bisphosphatase.	8.00E-19
AACW02000034_|_CONTIG_34__|_Rhizopus_oryzae_supercontig_3.1_|_[4377654-4477405]_|_99752_nt_	gene_12	0.70	0.91	gnl|CDD|30638 COG0290, InfC, Translation initiation factor 3 (IF-3).	3.00E-15

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000004_|_CONTIG_4__|_Rhizopus_oryzae_supercontig_3.1_|_[757263-1437475]_|_680213_nt_	gene_89	0.85	0.94	gnl|CDD|69205 pfam05669, SOH1, SOH1. The family consists of Saccharomyces.	6.00E-27
AACW02000041_|_CONTIG_41__|_Rhizopus_oryzae_supercontig_3.1_|_[5061334-5735491]_|_674158_nt_	gene_181	0.59	0.96	gnl|CDD|69362 pfam05832, DUF846, Eukaryotic protein of unknown function (DUF846).	2.00E-42
AACW02000042_|_CONTIG_42__|_Rhizopus_oryzae_supercontig_3.2_|_[1-107275]_|_107275_nt_	gene_11	0.38	0.92	gnl|CDD|68989 pfam05439, JTB, Jumping translocation breakpoint protein (JTB).	2.00E-06
AACW02000056_|_CONTIG_56__|_Rhizopus_oryzae_supercontig_3.2_|_[1578995-2467034]_|_888040_nt_	gene_317	0.76	0.79	gnl|CDD|69216 pfam05680, ATP-synt_E, ATP synthase E chain.	5.00E-10
AACW02000006_|_CONTIG_6__|_Rhizopus_oryzae_supercontig_3.1_|_[1509808-1953412]_|_443605_nt_	gene_42	0.96	1.00	gnl|CDD|68587 pfam05018, DUF667, Protein of unknown function (DUF667).	8.00E-73
AACW02000062_|_CONTIG_62__|_Rhizopus_oryzae_supercontig_3.2_|_[3248533-3874699]_|_626167_nt_	gene_132	0.98	1.00	gnl|CDD|32279 COG2096, COG2096, Uncharacterized conserved protein.	5.00E-44

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000062_|_CONTIG_62__|_Rhizopus_oryzae_supercontig_3.2_|_[3248533-3874699]_|_626167_nt_	gene_158	0.86	1.00	gnl|CDD|72839 pfam08583, UPF0287, Uncharacterised protein family (UPF0287).	3.00E-12
AACW02000073_|_CONTIG_73__|_Rhizopus_oryzae_supercontig_3.3_|_[601433-1277658]_|_676226_nt_	gene_193	0.96	1.00	gnl|CDD|80120 pfam05721, PhyH, Phytanoyl-CoA dioxygenase (PhyH).	1.00E-11
AACW02000074_|_CONTIG_74__|_Rhizopus_oryzae_supercontig_3.3_|_[1277853-1721144]_|_443292_nt_	gene_59	0.21	1.00	gnl|CDD|29102 cd00162, RING, RING-finger (Really Interesting New Gene).	3.00E-07
AACW02000075_|_CONTIG_75__|_Rhizopus_oryzae_supercontig_3.3_|_[1721245-2450281]_|_729037_nt_	gene_3	0.21	0.94	gnl|CDD|71421 pfam07985, SRR1, SRR1.	3.00E-08
AACW02000078_|_CONTIG_78__|_Rhizopus_oryzae_supercontig_3.3_|_[2477588-2558442]_|_80855_nt_	gene_36	1.00	1.00	gnl|CDD|67669 pfam04062, P21-Arc, P21-ARC (ARP2/3 complex 21 kDa subunit).	2.00E-56
AACW02000082_|_CONTIG_82__|_Rhizopus_oryzae_supercontig_3.3_|_[2888016-2926936]_|_38921_nt_	gene_3	1.00	1.00	gnl|CDD|67291 pfam03665, UPF0172, Uncharacterised protein family (UPF0172).	1.00E-37

Table A2 Continued

Chromosome/Contig_species	Predicted Gene_ID	Fraction of Aligned Residues		Domain Description	Alignment E-value
AACW02000083_|_CONTIG_83__|_Rhizopus_oryzae_supercontig_3.3_|_[2928881-3133860]_|_204980_nt_	gene_45	0.87	0.61	gnl|CDD|34689 COG5085, COG5085, Predicted membrane protein.	6.00E-12
AACW02000009_|_CONTIG_9__|_Rhizopus_oryzae_supercontig_3.1_|_[2111367-2400882]_|_289516_nt_	gene_99	0.47	0.96	gnl|CDD|32314 COG2131, ComEB, Deoxycytidylate deaminase.	1.00E-34
AACW02000091_|_CONTIG_91__|_Rhizopus_oryzae_supercontig_3.4_|_[1507490-2469713]_|_962224_nt_	gene_143	0.81	0.95	gnl|CDD|71948 pfam08520, DUF1748, Fungal protein of unknown function (DUF1748).	5.00E-18
AACW02000091_|_CONTIG_91__|_Rhizopus_oryzae_supercontig_3.4_|_[1507490-2469713]_|_962224_nt_	gene_191	0.47	0.82	gnl|CDD|58521 cd01846, fatty_acyltransferase_like.	2.00E-23

REFERENCES

Adams, M.D. S.E. Celniker et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287:** 2185-2195.

Alexandersson, M., S. Cawley et al. 2003. SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res* **13:** 496-502.

Allen, J.E., M. Pertea et al. 2004. Computational gene prediction using multiple sources of evidence. *Genome Res* **14:** 142-148.

Altschul, S.F., W. Gish et al. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Arumugam, M., C. Wei et al. 2006. Pairagon+N-SCAN_EST: a model-based gene annotation pipeline. *Genome Biol* **7 Suppl 1:** S5 1-10.

Ashurst, J.L., C.K. Chen et al. 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res* **33:** D459-465.

Atambayeva. 2008. Intron and exon length variation in Arabidopsis , rice, nematode, and human. *Molekulyarnaya Biologiya* **42:** 352.

Audic, S. and J.M. Claverie. 1998. Self-identification of protein-coding regions in microbial genomes. *Proc Natl Acad Sci U S A* **95:** 10026-10031.

Bafna, V. and D.H. Huson. 2000. The conserved exon method for gene finding. *Proc Int Conf Intell Syst Mol Biol* **8:** 3-12.

Balakirev, E.S. and F.J. Ayala. 2003. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* **37:** 123-151.

Baldi, P. 2000. On the convergence of a clustering algorithm for protein-coding regions in microbial genomes. *Bioinformatics* **16:** 367-371.

Barth, G. and C. Gaillardin. 1997. Physiology and genetics of the dimorphic fungus Yarrowia lipolytica. *FEMS Microbiol Rev* **19:** 219-237.

Batzoglou, S., L. Pachter et al. 2000. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res* **10:** 950-958.

Benson, D.A., I. Karsch-Mizrachi et al. 2008. GenBank. *Nucleic Acids Res* **36:** D25-30.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573-580.

Berget, S.M. 1995. Exon recognition in vertebrate splicing. *J Biol Chem* **270:** 2411-2414.

Berman, H.M., J. Westbrook et al. 2000. The Protein Data Bank. *Nucleic Acids Res* **28:** 235-242.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* **241:** 3-17.

Bernal, A., K. Crammer et al. 2007. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput Biol* **3:** e54.

Besemer, J. and M. Borodovsky. 1999. Heuristic approach to deriving models for gene finding. *Nucleic Acids Res* **27:** 3911-3920.

Besemer, J., A. Lomsadze et al. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29:** 2607-2618.

Birney, E., M. Clamp et al. 2004. GeneWise and Genomewise. *Genome Res* **14:** 988-995.

Bonaldo, M.F., G. Lennon et al. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* **6:** 791-806.

Boothroyd, J.C. and M.E. Grigg. 2002. Population biology of Toxoplasma gondii and its relevance to human infection: do different strains cause different disease? *Curr Opin Microbiol* **5:** 438-442.

Borodovsky, M., Y.A. Sprizhitsky, E.I. Golovanov and A.A. Alexandrov. 1986a. Statistical features in the Escherichia coli genome functional primary structure. II. Non-homogeneous Markov chains. *Molekuliarnaia biologiia* **20:** 833-840.

Borodovsky, M., Y.A. Sprizhitsky, E.I. Golovanov and A.A. Alexandrov. 1986b. Statistical features in the Escherichia coli genome functional primary structure. III. Computer recognition of protein coding regions. *Molekuliarnaia biologiia* **20**.

Borodovsky, M.Y. and J.D. McIninch. 1993. GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem.* **17:** 123-153.

Bray, N., I. Dubchak et al. 2003. AVID: A global alignment program. *Genome Res* **13:** 97-102.

Brejova, B., D.G. Brown et al. 2005. ExonHunter: a comprehensive approach to gene finding. *Bioinformatics* **21 Suppl 1:** i57-65.

Brudno, M., M. Chapman et al. 2003. Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics* **4:** 66.

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78-94.

Burset, M. and R. Guigo. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353-367.

Camargo, A.A., H.P. Samaia et al. 2001. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc Natl Acad Sci U S A* **98:** 12103-12108.

Carvalho, A.B., B.A. Dobo et al. 2001. Identification of five new genes on the Y chromosome of Drosophila melanogaster. *Proc Natl Acad Sci U S A* **98:** 13225-13230.

Charlesworth, B. 1996. The evolution of chromosomal sex determination and dosage compensation. *Curr Biol* **6:** 149-162.

Conrad, R., K. Lea et al. 1995. SL1 trans-splicing specified by AU-rich synthetic RNA inserted at the 5' end of Caenorhabditis elegans pre-mRNA. *RNA* **1:** 164-170.

Consortium, T.C.e.S. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282:** 2012-2018.

Cuny, G., P. Soriano et al. 1981. The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. *Eur J Biochem* **115:** 227-233.

Cuomo, C.A., U. Guldener et al. 2007. The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317:** 1400-1402.

Curwen, V., E. Eyras et al. 2004. The Ensembl automatic gene annotation system. *Genome Res* **14:** 942-950.

Dean, R.A., N.J. Talbot et al. 2005. The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature* **434:** 980-986.

Dehal, P., Y. Satou et al. 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science* **298:** 2157-2167.

Delcher, A.L., D. Harmon et al. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27:** 4636-4641.

Delcher, A.L., S. Kasif et al. 1999. Alignment of whole genomes. *Nucleic Acids Res* **27:** 2369-2376.

Delcher, A.L., A. Phillippy et al. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30:** 2478-2483.

Dempster, A.P., N.M. Laird et al. 1977. Maximum likelihood from inclomplete data via the EM algorithm. *J.R.Statis. Soc.***:** 39:31--38.

Dietrich, F.S., S. Voegeli et al. 2004. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science* **304:** 304-307.

Dressman, D., H. Yan et al. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A* **100:** 8817-8822.

Drutz, D.J. and A. Catanzaro. 1978. Coccidioidomycosis. Part I. *Am Rev Respir Dis* **117:** 559-585.

Dujon, B., D. Sherman et al. 2004. Genome evolution in yeasts. *Nature* **430:** 35-44.

Durbin, R., Eddy, S, Krogh A, Mitchison G. 1998. *Biological sequence analysis*. Cambridge University Press, Cambridge.

Durkin, M., S. Kohler et al. 2001. Chronic infection and reactivation in a pulmonary challenge model of histoplasmosis. *J Infect Dis* **183:** 1822-1824.

Fairbrother, W.G., R.F. Yeh et al. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297:** 1007-1013.

Foissac, S., P. Bardou, A. Moisan, et al. 2003. EUGÈNE'HOM: a generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Research* **31:** 3742-3745.

Frishman, D., A. Mironov et al. 1998. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* **26:** 2941-2947.

Frishman, D., Mironov, A., Mewes, H. W., Gelfand, M. 1998. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* **26:** 2941-2947.

Galagan, J.E., S.E. Calvo et al. 2003. The genome sequence of the filamentous fungus Neurospora crassa. *Nature* **422:** 859-868.

Galagan, J.E., S.E. Calvo et al. 2005. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature* **438:** 1105-1115.

Gertz, E.M., Y.K. Yu et al. 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol* **4:** 41.

Ghadessy, F.J., J.L. Ong et al. 2001. Directed evolution of polymerase function by compartmentalized self-replication. *Proc Natl Acad Sci U S A* **98:** 4552-4557.

Gish, W. and D.J. States. 1993. Identification of protein coding regions by database similarity search. *Nat Genet* **3:** 266-272.

Goffeau, A.e.a. 1997. The yeast genome directory. *Nature* **387:** 5.

Guigo, R. 1998. Assembling genes from predicted exons in linear time with dynamic programming. *J Comput Biol* **5:** 681-702.

Guigo, R., P. Flicek et al. 2006. EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol* **7 Suppl 1:** S2 1-31.

Guigo, R., S. Knudsen et al. 1992. Prediction of gene structure. *J Mol Biol* **226:** 141-157.

Hayes, W.S. and M. Borodovsky. 1998. Deriving ribosomal binding site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction. *Pac Symp Biocomput***:** 279-290.

Hebsgaard, S.M., P.G. Korning et al. 1996. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res* **24:** 3439-3452.

Holt, R.A. G.M. Subramanian et al. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298:** 129-149.

Howe, K.L., T. Chothia et al. 2002. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res* **12:** 1418-1427.

Hurley, J.H. 2008. ESCRT complexes and the biogenesis of multivesicular bodies. *Curr Opin Cell Biol* **20:** 4-11.

Initiative, T.A.G. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815.

Initiative:, T.A.G. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815.

Katti, M.V., P.K. Ranjekar et al. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol Biol Evol* **18:** 1161-1167.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Kent, W.J. and A.M. Zahler. 2000. Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. *Genome Res* **10:** 1115-1125.

Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5:** 59.

Korf, I., P. Flicek et al. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1:** S140-148.

Krizman, D.B., R.F. Chuaqui et al. 1996. Construction of a representative cDNA library from prostatic intraepithelial neoplasia. *Cancer Res* **56:** 5380-5383.

Krogh, A. 2000. Using database matches with for HMMGene for automated gene detection in Drosophila. *Genome Res* **10:** 523-528.

Krogh, A., M. Brown et al. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235:** 1501-1531.

Kullback, S., and Leibler, R. A., . 1951. On information and sufficiency. *Annals of Mathematical Statistics* **22:** 79-86.

Kulp, D., D. Haussler et al. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol* **4:** 134-142.

Kurtz, S., A. Phillippy et al. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5:** R12.

Lafferty, J., McCallum,A., Pereira,F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Eighteenth International Conference on Machine Learning (ICML-2001).*

Larsen, T.S. and A. Krogh. 2003. EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* **4:** 21.

Laub, M. and D. Smith. 1998. Finding intron/exon splice junctions using INFO, INterruption Finder and Organizer. *J Comput Biol.* **5:** 307-321.

Lawrence, C.E., S.F. Altschul et al. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262:** 208-214.

Lim, L.P. and C.B. Burge. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* **98:** 11193-11198.

Liolios, K., N. Tavernarakis et al. 2006. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res* **34:** D332-334.

Liu, Q., A.J. Mackey et al. 2008. Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics* **24:** 597-605.

Loftus, B.J., E. Fung et al. 2005. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science* **307:** 1321-1324.

Lomsadze, A., V. Ter-Hovhannisyan et al. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* **33:** 6494-6506.

Lukashin, A.V. and M. Borodovsky. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26:** 1107-1115.

Lukashin, A.V., J. Engelbrecht et al. 1992. Multiple alignment using simulated annealing: branch point definition in human mRNA splicing. *Nucleic Acids Res* **20:** 2511-2516.

Majoros, W.H., M. Pertea et al. 2003. GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders. *Nucleic Acids Res* **31:** 3601-3604.

Margulies, M., M. Egholm et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Mathe, C., M.F. Sagot et al. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res* **30:** 4103-4117.

Merchant, S.S. S.E. Prochnik et al. 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318:** 245-250.

Meyer, I.M. and R. Durbin. 2002. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics* **18:** 1309-1318.

Mironov, A.A., M.A. Roytberg et al. 1998. Performance-guarantee gene predictions via spliced alignment. *Genomics* **51:** 332-339.

Mitchell, T.G. and J.R. Perfect. 1995. Cryptococcosis in the era of AIDS--100 years after the discovery of Cryptococcus neoformans. *Clin Microbiol Rev* **8:** 515-548.

Mitrophanov, A.Y., Lomsadze, A. and Borodovsky, M. 2005. Sensitivity of hidden Markov models. *J. Appl. Probab.* **42:** 632-642.

Morgenstern, B. 2000. A space-efficient algorithm for aligning large genomic sequences. *Bioinformatics* **16:** 948-949.

Morgenstern, B., A. Dress et al. 1996. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc Natl Acad Sci U S A* **93:** 12098-12103.

Neverov, A.D., M. S. Gelfand, and A. A. Mironov. 2003. GipsyGene: A Statistics-Based Gene Recognizer for Fungal Genomes. *Biophysics* **Vol. 48:** S71–S75.

Notredame, C. and K. Suhre. 2004. Computing multiple sequence/structure alignments with the T-coffee package. *Curr Protoc Bioinformatics* **Chapter 3:** Unit3 8.

Orr, H.A. and Y. Kim. 1998. An adaptive hypothesis for the evolution of the Y chromosome. *Genetics* **150:** 1693-1698.

Pachter, L., S. Batzoglou et al. 1999. A dictionary-based approach for gene annotation. *J Comput Biol* **6:** 419-430.

Parra, G., P. Agarwal et al. 2003. Comparative gene prediction in human and mouse. *Genome Res* **13:** 108-117.

Parra, G., E. Blanco et al. 2000. GeneID in Drosophila. *Genome Res* **10:** 511-515.

Parra, G., K. Bradnam et al. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23:** 1061-1067.

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D. V., Leroy, P., Rouze, P. 1999. Evaluation of gene prediction software using a genomic data set: application to Arabidopsis thaliana sequences. *Bioinformatics* **15:** 887-899.

Pedersen, J.S. and J. Hein. 2003. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics* **19:** 219-227.

Pel, H.J., J.H. de Winde et al. 2007. Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. *Nat Biotechnol* **25:** 221-231.

Pertea, M., X. Lin et al. 2001. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res* **29:** 1185-1190.

Peterson, L.A., M.R. Brown et al. 1998. An improved method for construction of directionally cloned cDNA libraries from microdissected cells. *Cancer Res* **58:** 5326-5328.

Pimpinelli, S., M. Berloco et al. 1995. Transposable elements are stable structural components of Drosophila melanogaster heterochromatin. *Proc Natl Acad Sci U S A* **92:** 3804-3808.

Piotrowska, M., R. Natorff et al. 2000. sconC, a gene involved in the regulation of sulphur metabolism in Aspergillus nidulans, belongs to the SKP1 gene family. *Mol Gen Genet* **264:** 276-282.

Pruitt, K.D., T. Tatusova et al. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35:** D61-65.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE* **77:** 257-286.

Raetsch, G.a.S., S. 2005. *Large Scale Semi Hidden Markov SVMs. In Advances in Neural Information Processing Systems*. Cambridge, MA.

Redecker, D., R. Kodner et al. 2000. Glomalean fungi from the Ordovician. *Science* **289:** 1920-1921.

Reese, M.G., D. Kulp et al. 2000. Genie--gene finding in Drosophila melanogaster. *Genome Res* **10:** 529-538.

Reese, M.G., Kulp, D., Tammana, H., Haussler, D. 2000. Genie--gene finding in Drosophila melanogaster. *Genome Res* **10:** 529-538.

Reinke, V., H.E. Smith et al. 2000. A global profile of germline gene expression in C. elegans. *Mol Cell* **6:** 605-616.

Rinner, O. and B. Morgenstern. 2002. AGenDA: gene prediction by comparative sequence analysis. *In Silico Biol* **2:** 195-205.

Rogic, S., B.F. Ouellette et al. 2002. Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics* **18:** 1034-1045.

Rogozin, I.B., L. Milanesi et al. 1996. Gene structure prediction using information on homologous protein sequence. *Comput Appl Biosci* **12:** 161-170.

Rollenhagen, C., C.A. Hodge et al. 2004. The nuclear pore complex and the DEAD box protein Rat8p/Dbp5p have nonessential features which appear to facilitate mRNA export following heat shock. *Mol Cell Biol* **24:** 4869-4879.

Salamov, A.A. and V.V. Solovyev. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10:** 516-522.

Salzberg, S.L., M. Pertea et al. 1999. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59:** 24-31.

Sarawagi, S., Cohen, W. 2004. Semi-Markov Conditional Random Fields for Information Extraction. In *Advances in Neural Information Processing Systems, NIPS* Vancouver, British Columbia, Canada.

Schiex, T., A. Moisan and P. Rouze. 2001. EuGene: an eucaryotic gene finder that combines several sources of evidence. *Computational biology*.

Schneider, T.D. and R.M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18:** 6097-6100.

Schreiber, G. 2002. The evolutionary and integrative roles of transthyretin in thyroid hormone homeostasis. *J Endocrinol* **175:** 61-73.

Schwartz, S., L. Elnitski et al. 2003. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res* **31:** 3518-3524.

Schwartz, S., Z. Zhang et al. 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10:** 577-586.

Scott, D.C. and R. Schekman. 2008. Role of Sec61p in the ER-associated degradation of short-lived transmembrane proteins. *J Cell Biol* **181:** 1095-1105.

Siepel, A. and D. Haussler. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol* **11:** 413-428.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. NY Chapman and Hall

Skaletsky, H., T. Kuroda-Kawaguchi et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825-837.

Sonnenburg, S., Raetsch, G., and Schoelkopf, B. 2005. Large Scale Genomic Sequence SVM Classifiers. In *22nd International Machine Learning Conference*. ACM Press.

Sparks M., B.M., and Dorman K. 2007. *Bioinformatics Research and Applications*. Springer Berlin Heidelberg, Berlin.

Stanke, M., O. Schoffmann et al. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7:** 62.

Stanke, M. and S. Waack. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2:** II215-II225.

Tavare, S. and B. Song. 1989. Codon preference and primary sequence structure in protein-coding regions. *Bull Math Biol* **51:** 95-115.

Taylor, J.W. and M.L. Berbee. 2006. Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia* **98:** 838-849.

Ter-Hovhannisyan, V., A. Lomsadze et al. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res*.

Thompson, W., E.C. Rouchka et al. 2003. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* **31:** 3580-3585.

Vicoso, B. and B. Charlesworth. 2006. Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet* **7:** 645-653.

Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory* **13:** 260–269.

Wang, Z. and C.B. Burge. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14:** 802-813.

Watanabe, K. and S. Harayama. 2001. [SWISS-PROT: the curated protein sequence database on Internet]. *Protein, Nucleic Acid and Enzyme* **46:** 80-86.

Wood, V. R. Gwilliam et al. 2002. The genome sequence of Schizosaccharomyces pombe. *Nature* **415:** 871-880.

Yeh, R.F., L.P. Lim et al. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res* **11:** 803-816.

Yu, J., S. Hu et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296:** 79-92.